Inferential analysis of genomic 3D organization

Jonas Paulsen

Thesis for the degree of Philosophiae Doctor (PhD)

Institute for Cancer Genetics and Informatics
The Norwegian Radium Hospital, Oslo University Hospital
Faculty of Medicine
University of Oslo, Norway
2014

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 1991

ISBN 978-82-8264-950-6

Cover: Hanne Baadsgaard Utigard.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Akademika Publishing.
The thesis is produced by Akademika Publishing merely in connection with the thesis defence. Kindly direct all inquiries regarding the thesis to the copyright holder or the unit which grants the doctorate.

Acknowledgements

The work included in this thesis was funded by the Institute for Cancer Genetics and Informatics at the Oslo University Hospital, and was supervised by Prof. Eivind Hovig at the Department of Tumor Biology at Oslo University Hospital. During the entire PhD-period, from March 2011 until May 2014, I have been lucky to have a nice office space in the new research building at the Department of Informatics at the University of Oslo.

First, and foremost, I would like to thank Prof. Eivind Hovig for providing me with the perfect balance of independency and guidance in his supervision during the project. His eagerness to see results, while still allowing me the time to learn from mistakes, has been truly valuable. Due to his large network of colleagues, I have been lucky to collaborate with a lot of people during this project. Looking back, it is clear that my project has been very cross-disciplinary, with collaborators from statistics, informatics and the medical sciences. I am grateful to all the people who have been part of my academic life during the last years, and for all the enthusiasm and lively discussions. I would particularly like to thank Tonje G. Lien for interesting and thorough discussions particularly related to the first paper, where we collaborated closely. Special thanks also go to co-supervisor Prof. Geir Kjetil Sandve, who introduced me to the Genomic HyperBrowser, and who has actively and enthusiastically helped out with many of the issues arising along the way. The same goes for Dr. Sveinung Gundersen, who has also contributed greatly with his GTrack file format system. I would also like to thank co-supervisor Prof. Arnoldo Frigessi, who has been of great help particularly during the start of the project, and who is always very lively and engaged. Similarly, I would like to thank Prof. Ingrid Glad, Dr. Lars Holden, Dr. Marit Holden, Prof. Ørnulf Borgan and Dr. Einar Rødland for eagerly helping out with important contributions to the statistics. Thanks also to Tobias G. Waaler for contributions to some of the functionality underlying HiBrowse.

I am also very grateful for the warm and welcoming environment at the bioinformatics group at Institute for Informatics. Particularly, I would like to thank my office buddy Dr. Sigve Nakken for being so nice and friendly, and patiently answering all my PhD-related questions.

Finally, I would like to give a warm thank you to my family, who has been very supportive and understanding during the ups and downs of the entire PhD-period. Particularly, I am grateful to Solveig, my soon-to-be wife, whose extreme patience and care for others have allowed me to pursue my dreams to the fullest.

Oslo, May 2014
Jonas Paulsen

Contents

Glossary and List of Abbreviations

3C chromosome conformation capture.

3D three-dimensional.

4C chromosome conformation capture-on-chip (or circular chromosome conformation capture).

5C chromosome conformation capture carbon copy.

anchor (ChIA-PET) genomic region, identified in the ChIA-PET procedure, where a protein of choice binds to DNA and where 3D interactions can occur.

BED browser extensible data. Simple format for representation of genomic positions with accompanying annotation.

bin (Hi-C) a fixed-length segment of DNA encompassing a set of restriction fragments for which interaction frequencies are aggregated and quantified in a final contact matrix.

bp base pair (of DNA).

CAP chromatin architectural protein.

CCD conserved consecutive distances.

ChIA-PET chromatin interaction analysis with paired-end tag sequencing.

ChIP chromatin immunoprecipitation.

contact frequency see interaction frequency.

CpG a site in DNA where a cytosine and a guanine occur next to each other.

CTCF CCCTC-binding factor.

DHS DNAse I hypersensitive site.

DNA deoxyribonucleic acid.

FDR false discovery rate.

FISH fluorescence in situ hybridization.

GC-content the amount of guanine (G) and cytosine (C) in a segment of DNA, often represented as the percentage of these bases compared to all bases in the given segment.

GCC genome conformation capture.

genomic distance the distance, as measured in number of base pairs, bins or segments, along the linear sequence of bases in DNA.

GUI graphical user interface.

interaction frequency the number of times a ligation product is detected between two genomic regions, resulting from physical proximity between the two regions.

interchromosomal between chromosomes.

intrachromosomal within chromosomes.

kb kilo base, 1000 base pairs of DNA.

LAD lamina associated domain.

LCR locus control region.

LGP linked genome partition.

lncRNA long non-coding RNA.

LP linked point.

LS linked segment.

LVP linked valued point.

LVS linked valued segment.

Mb mega base, 1 million base pairs of DNA.

MC Monte Carlo.

NAD nucleolus-associated chromatin domain.

NCHG non-central hypergeometric.

nm nanometer.

P point.

PCA principal component analysis.

PcG Polycomb group.

PCR polymerase chain reaction.

qPCR quantitative polymerase chain reaction.

RNA ribonucleic acid.

RNAP II RNA polymerase II.

rRNA ribosomal RNA.

S segment.

sequence-based distance see genomic distance.

SNP single nucleotide polymorphism.

TAD topologically associating domain.

TCC tethered chromosome conformation capture.

TF transcription factor.

TSS transcription start site.

XCI X-chromosome inactivation.

Z-score standard score, the (positive or negative) number of standard deviations (sd) an observation (obs) is from the mean. Calculated as [obs-mean]/sd.

List of Papers

Paper I Paulsen J, Lien TG, Sandve GK, et al. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. Nucleic acids research 2013;41:5164–5174

Paper II Paulsen J, Sandve GK, Gundersen S, Lien TG, Trengereid K, and Hovig E. Hi-Browse: multi-purpose statistical analysis of genome-wide chromatin 3D organization. Bioinformatics 2014;30:1620–1622

Paper III Paulsen J, Rødland EA, Holden L, Holden M, and Hovig E. A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. (Manuscript in review)

Chapter 1

Introduction

The study of the structure and function of DNA and chromatin goes back several centuries. The interest to understand the mechanisms by which characteristics where inherited from one generation to the next was boosted dramatically with the publishing of Darwin's theory of evolution by natural selection [1]. Remarkably, at the same time as Darwin's famous publication, the austrian scientist, and friar, Gregor Mendel, conducted experiments on pea plants showing that characteristics were inherited according to particular mathematical rules. Unfortunately, Darwin was never aware of Mendel's pioneering work, which was largely ignored until its rediscovery in the beginning of the 1900s [2]. As most of the fundamental insights into the mechanisms of heredity was mapped out during the first part of the 1900s, the functioning of the DNA and the genes became the central focus of the second half of the century. The finishing of the sequence of the human genome at the start of the 21st century [3], resulted in an explosion in new technologies for mapping out functional and regulatory mechanisms of cells and tissues that can be linked to the underlying sequence. The result has been a shift from understanding single genes, to a more general approach where the entire genome has become the system of study (genomics). Recently, with large-scale projects such as the ENCODE project [4] and the Epigenome Roadmap [5], insights into the epigenomic regulation of various tissues and cell-lines are starting to emerge.

The 6 billion bases that constitute the entire diploid human genome make up a total of about 2 meters of DNA inside each cell. Considering that the diameter of the nucleus is typically around 10-20 micrometers, the chromatin fibre needs to be compacted and folded to an extreme degree [6]. Novel technologies for mapping genome-wide 3D interactions between distal regions in the genome have, during the last years, allowed for probing this structure for the very first time. Due to massive improvements in throughput of such methods, ever-increasing amounts of data are being produced.

The topic of this thesis is the statistical and computational analysis of data from such methods, with particular focus on inferential analysis. In the introductory part of the thesis, a brief review of the history of chromatin biology will be given, in addition to a summary of the major insights that has been gained in recent years. After this, an introduction to some of the technologies for mapping genomic 3D interactions will be given, and finally the computational and statistical methods for analyzing such data will be reviewed. The introductions given will mostly focus on mammalian systems, particularly human data, since this is also the focus of the thesis.

1.1 Chromatin biology

1.1.1 A historical perspective

It was the German biologist Ernst Haeckel who first proposed the idea that the nucleus takes care of inheritance in eukaryotic cells, in his now famous book Generelle Morphologie from 1866 [7]. However, the understanding that the nucleus contained nucleic acids (DNA), or "nuclein" as it was initially called, was first proposed in 1871 by Miescher [8]. In the early 1880s this led Walther Flemming, and independently Edouard Van Beneden, to zoom in on the structures within the nucleus. Using aniline dyes, Flemming was able to visualize and describe in details the structures within the nucleus, naming them "chromatin" [9, 10]. The link between nuclein and chromatin, and subsequently genetic inheritance was beginning to emerge [11].

While the substance of inheritance, in the form of chromatin, was understood to be found within the nucleus, the units of inheritance was not described until the early 1900s, when Theodor Boveri proposed that chromosomes were fundamental for embryonic development and inheritance [12]. At the same time, Walter Sutton became one of the first to couple Mendel's heredity rules to the chromosomes themselves, giving a convincing argument for the chromosome theory of heredity, known today as the Boveri-Sutton chromosome theory [13].

It was understood at the time that chromatin was consisting of a mixture of both protein (histones) and nucleic acid (DNA), but little was known about which of these substances that was most important for genetic inheritance. However, in the 1940s, the idea that it was DNA that formed the ultimate basis of inheritance was starting to emerge [14]. The central role of DNA was later confirmed in 1952 in a famous experiment on the T2 phage, by Alfred Hershey and Martha Chase [15]. A year later, James D. Watson and Francis Crick proposed a model for the double-helical structure of DNA [16]. Crick later proposed the central dogma of biology (DNA makes RNA makes proteins), and proposed that protein-coding DNA was made up of non-overlapping codon triplets, which led to the deciphering of the genetic code [17]. These findings laid out the foundations of genetics, and spawned novel insights in the last decades of the 20th century.

In parallel, further studies on the structure of chromatin and chromosomes were being conducted. While it was known at the time that histones could be modified by acetylation and methylation, very little was known about the function of such modifications. However, in 1964, Allfrey, Faulkner and Mirsky proposed for the first time that such histone modifications were related to gene regulation and expression [18]. These findings suggested that modifications of histones could regulate the transcription of individual genes along chromosomes.

The interest in studying the details of the chromatin fibre itself was also boosted in the early 1960s. Particularly, the technique of X-ray diffraction imaging that had been so successful in determining fundamental structural properties of protein [19, 20] and DNA [21] was used on chromatin in the hope to elucidate the underlying structural properties. This led to the superhelix model of chromatin [22], where the DNA double helix was thought to be further coiled into a helical chromatin fiber superstructure, which was believed to be stabilized by histone interactions.

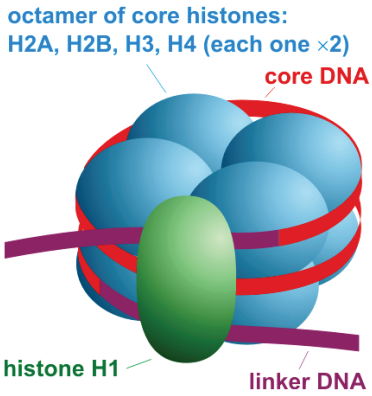

Figure 1.1: Schematic illustration of the organization of a nucleosome. The nucleosome consists of an octamer of histone pairs (H2A, H2B, H3 and H4). In addition, the histone H1 protein binds to the linker DNA to stabilize the nucleosome. The DNA string is wrapped around the nucleosome (as illustrated) by approximately 147 base pairs of DNA. Image source: Wikimedia Commons

These views were drastically undermined in the beginning of the 1970s with the nucleosome model, stating that DNA was wrapped around histone octamers, with coils of around 200 base pairs [23] (see Figure 1.1). The resulting model of chromatin became known as the "beads-on-a-string" model, and electron micrographs of chromatin fibres showed clear evidence that such structures existed [24]. This revolutionary insight was immediately recognized as a fundamental property of chromatin and gene regulation, since the DNA would be accessible to binding of other proteins, while the histones would provide an easy way to re-package the DNA based on their chemical properties.

These insights led Finch and Klug to propose a new higher-order structural model of chromatin deemed to replace the superhelix model. This new model became known as the 30-nm chromatin fibre model [25]. In this model, nucleosomes are packaged in a solenoid super-structure where consecutive nucleosomes are adjacent to each other, forming a helical structure. This model was quickly accepted, since electron microscopy studies at the time supported the view of chromatin as a 30 nanometer fibre [24].

With the advent of fluorescence in situ hybridization (FISH) in the early 1980s, visualization of the spatial positioning of chromosomal regions became possible. With FISH, hybridization of specific probes with fluorescent dyes could be used to visualize specific DNA regions within the nucleus [26] (see section 1.2.1). These studies confirmed earlier views that chromosomes were not randomly positioned within the nucleus [27, 28]. Of particular importance was the visualization of the relative positioning of entire chromosomes, and the discovery that individual chromosomes seemed to occupy distinct parts of the nucleus, forming so-called "chromosome territories" [29].

1.1.2 A modern view of the structure of chromatin

In the post-genomic era, the focus has shifted from understanding the physical properties of chromatin, to a more unified view of how physical and functional properties combine to determine regulatory roles in gene expression and cell differentiation. A major breakthrough

facilitating such studies came with the chromosome conformation capture (3C) technique that allows for identification of physical interactions linked directly to the underlying sequence. Many of the resulting discoveries have confirmed the earlier view of chromatin as a dynamic, yet conserved, structure. However, the emerging view of the physical properties of the chromatin fibre is more complex than the 30-nm chromatin fibre of Finch and Klug. It has instead been proposed that the state of the chromatin fibre in living interphase nuclei resembles a dynamic, fractal-like polymer, which is capable of undergoing dynamic and rapid changes without a regular fibre-structure above the 10 nanometer scale [30, 31].

1.1.2.1 The chromosome

One of the most prominent features of eukaryotic chromosomes is the highly orchestrated way in which they are repositioned and packaged during cell division [9]. It was early recognized that the centromeres have a special function in chromosomal architecture, since they serve the function of linking sister chromatids together during mitosis, via the kinetochores. Also of great importance are the repetitive sequences at the ends of each of the chromatids, called telomeres, since they protect the chromosomes from destroying genes during end-degradation during chromosome replication [32]. In Figure 1.2, an overview of some of the large-scale attributes of eukaryotic chromosomes is given.

The parts of the chromosomes that serve structural functions, such as centromeres and telomeres are largely composed of what is called constitutive heterochromatin, a highly compact chromatin structure often seen at repetitive regions. Heterochromatin was discovered due to the fact that compact chromatin was more deeply stained [33], and was speculated to be gene-poor and largely inactive (“inert”) [34]. Facultative heterochromatin on the other hand, has been shown to be a much more dynamic structure that can switch from the compact state to a more open and active state, for example during cell differentiation. Euchromatin constitutes the bulk of the genome of humans and other mammals and is, in contrast to heterochromatin, mostly rich in genes and often undergoing active transcription. [35].

The precise structural differences between heterochromatin and euchromatin is largely not understood, even though it is assumed that heterochromatin is defined by a more extensive looped structure compared to more open regions of the genome that resemble the beads-on-a-string model. The open chromatin structure allows the transcription machinery, as well as regulatory proteins, to bind more easily (“accessibility hypothesis”) [35].

1.1.2.2 Chromatin architecture

The more finely detailed regional differences along chromosomes were not discovered until the late 1960s, when differences in banding patterns due to differential staining affinities attributed to differences in GC-content of the chromosomes were observed [36]. Today, regional variation in GC-content is seen as a fundamental property of chromatin function, since several important regulatory roles are directly linked to the GC-content of the underlying DNA.

DNA methylation One such regulatory function is the covalent addition of a methyl-group at the cytosine in a CpG dinucleotide. Such covalent DNA methylation modifications at or

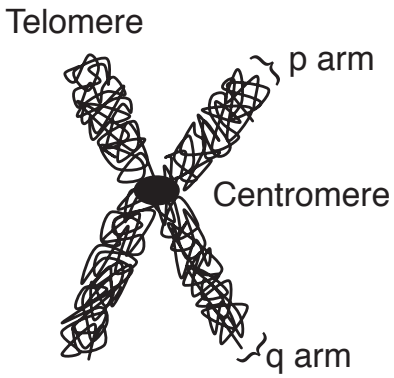

Figure 1.2: Schematic illustration of a condensed, replicated chromosome. Chromosome arms are named according to their relative length, with p indicating the shorter arm, while q indicates the longer arm of each chromatid. The telomeres are found as a repetitive sequence at the end of each of the four arms. The centromere, indicated as a black circle in the center, serves the function of linking sister chromatids together during mitosis.

around genes have been shown to cause transcriptional repression. Interestingly, since the methylated cytosine in CpG dinucleotides is prone to mutate into a thymine (to form a TpG dinucleotide), there is a depletion of CpG dinucleotides in mammalian genomes compared to what can be expected by pure chance. However, around genes that are required and therefore expressed in most tissues (housekeeping genes), this depletion is not seen, since CpG dinucleotides will almost never be methylated at these positions. Therefore, over evolutionary time these sites will be the only sites that do not gradually get depleted of CpG dinucleotides [35]. Such un-depleted regions are known as CpG islands, and are found at approximately 40% of the promoters of human genes [37].

Histone modifications The histone proteins are also subject to covalent modifications, such as methylation and acetylation, typically at their N-terminal tails. These modifications can change the charge of the overall chromatin structure, and thereby regulate the degree of compactness of the chain ("charge neutralization model"). Additionally, the combination of various histone modifications at a given site could together constitute a "histone code" that regulates genomic function directly. Finally, the histone modifications themselves could act as platforms that facilitate binding of enzymes that alter chromatin such that chromatin altering feedback loops are initiated ("signalling pathway model") [38].

Today, a large number of histone modifications in several tissues and cell-types have been mapped. Examples of known repressive modifications (in mammalian genomes) are the H3K9me3 (and H3K9me2) modifications that mark heterochromatin, and the H3K27me3 modification that marks Polycomb repressed regions (discussed in section 1.1.3.6). Active regulatory elements known as enhancers are, on the other hand, often marked by H3K4me1, or various histone acetylations. Active promoter regions, which are often found close to the transcription start sites (TSS) of genes, are often marked by H3K4me3. Interestingly, promoter regions can also be found in a poised state containing both inactivating and activating marks, such as DNA methylation, H3K27me3 and H3K4me3 simultaneously. These "biva-

lent" elements are especially abundant in embryonic stem cells, since many of the genes are ready to turn on transcription as the cells differentiate into different lineages. Inside the gene-body, a clear pattern of increasing H3K36me3, and decreasing H3K4me1 downstream into the gene is often seen. In Figure 1.3, an overview of the various mechanisms that characterizes the dynamics of chromatin is given [38]. The nucleosomes themselves are also highly dynamic, and can be unwrapped and repositioned around the histone octamer structure. For example, nucleosomes are depleted upstream of actively transcribed genes [39].

Insulators The extent of specific histone modifications, such as H3K27me3, at genomic sites, is itself being regulated. A common mechanism for such regulation is through so called insulator elements. These elements can block the spread of the highly compact structure of heterochromatin to move into active regions of chromatin, or prevent active regions to spread into inactive regions causing unwanted expression of genes. Such a chromatin barrier activity is thought to occur via recruitment of histone acetyltransferases that counter-act the spreading of repressive chromatin (see section 1.1.2.3). In mammals, the most studied insulator protein is the CCCTC-binding factor (CTCF) [35, 40] (see section 1.1.3.5, for details).

In addition to barrier activity, insulators have been shown to be able to block looping interactions between enhancers and promoters. This type of insulator function is assumed to allow regulated activity of cell-type specific gene expression in different tissues. The mechanisms of such enhancer-blocking activity have been largely unknown. Recently, however, it has been speculated that the activity of CTCF, in combination with cohesin, can achieve much of its regulatory function via facilitating looping and domain formation (see section 1.1.3.4 for details).

Chromatin states The emerging view of the architecture of chromatin is that of a dynamic structure with multiple functional states defined by the combinations of histone modifications, DNA methylation and protein binding patterns on the underlying chromatin [41–43]. Enhancer regions are particularly dynamic, and are also highly abundant with at least 400,000 candidate sites in the human genome. Differences in H3K4 methylation state, acetylation and DNA-methylation are all thought to cooperate in determining the activity state patterns of individual enhancers. Specifically, enhancers marked by both H3K27ac and H3K4me1 are typically active, while enhancers marked only by H3K4me1 are thought to be primed for activation. As mentioned previously, a bivalent (or poised) state consisting of a combination of H3K27me3 (repressive) and H3K4me1 (activating) marks is also seen, and is particularly present in embryonic stem cells [44].

The exact definition of chromatin states is not clear, as illustrated by the wide range of number of states identified for human data. For example, in a paper by Ernst et al., a total of 51 distinct chromatin states was found, based on a large set of histone modifications and protein binding patterns [42]. Using a similar method on fewer histone modifications across nine different cell-lines, the same authors discovered and described 15 different chromatin states [45]. Similar results have been seen when investigating chromatin states in *Drosophila melanogaster* as well. For example, when analyzing the combinatorial patterns of 53 different chromatin proteins, five chromatin types defined according to the underlying activity patterns of the *Drosophila* genome, was found [43]. In another study, by investigating a set of histone

modifications, nine chromatin types were found [46].

Recently, by mapping chromatin states consisting of five histone modifications in combination with CTCF and cohesin in 19 individuals, Kasowski et al. were able to show a remarkable variability in chromatin state patterns between the individuals [47]. The highest variability was found for active chromatin marks at enhancer states, and at repressive marks consisting of H3K27me3. Due to the high variability of chromatin states, and the uncertainties in the separation between chromatin states depending on which factors that are assessed, further research needs to be conducted before the complete nature of the chromatin state activity patterns along chromosomes is revealed [48].

1.1.2.3 Chromatin dynamics and regulation

The chromatin fibre is regulated and organized via a series of proteins and protein complexes that alter the various modifications on histones and on the DNA itself. Such proteins are known as chromatin architectural proteins (CAPs), and are characterized by their ability to recognize features on chromatin, such as DNA-methylation or specific histone modifications. The recognition and binding of the CAPs usually involves a structural change in chromatin, for example by creation of repressive and compacted chromatin. Specifically, heterochromatin protein 1 (HP1) binds to methylated histone H3K9, and thereby recruits a histone methyltransferase (HMT) enzyme, which further methylates H3K9 in adjacent nucleosomes. This spread of repressive histone marks (usually until an insulator element is reached) causes the chromatin to be tightly packed and inaccessible [49].

A similar mechanism is seen with the Polycomb group proteins (PcG), which recognize H3K27 and induce methylation. This is done by the PRC2 complex that catalyzes methylation on H3K27, which in turn is recognized by the PRC1 complex, and binds to the methylated histones. The PRC1 complex then catalyses further H3K27 methylation on adjacent nucleosomes. Interestingly, the methylated histone H3 will be partially inherited during cell division such that the PRC1 complex can maintain the repressive chromatin marks in the daughter cells as well. However, as differentiation proceeds and repressed genes are needed, the H3K27 methylation is lost and will no longer be maintained [35, 49].

Certain CAPs can also bind to methylated DNA, as for example methyl-CpG binding protein 2 (MeCP2). This protein binds specifically to methylated CpG, which causes a tightly packed and closed chromatin structure, and consequently transcriptional repression. Binding of MeCP2 results in the recruitment of histone deacetylase (HDAC) that induces the removal of (activating) acetyl groups from histones, further de-activating the chromatin structure [35].

There are also examples of CAPs that activate and open chromatin, such as histone acetyltransferases (HATs). These multi-protein complexes do the opposite job of HDACs, by acetylating lysine residues on histones, which reduces (neutralizes) the net positive charge on the histone molecules. Many proteins involved in activation of gene expression have been shown to induce HAT activity, and to be capable of acetylating histone molecules [35].

The mechanistic understanding on the effect of such chromatin changes is not complete. However, two different, yet overlapping, models exist of these effects. In the “direct” model, alterations in the chemical composition (acetylation, methylation etc.) of histone molecules change the compaction of the chromatin fibre, making the DNA accessible to the transcription

machinery, including transcription factors. In the "effector-mediated" model, the effector proteins are instead thought to "read" the shape and chemical composition of the histones and their modifications, and act accordingly by initiating downstream processes, such as recruiting other chromatin protein complexes [50].

The linker histone H1 itself is also of major importance for stabilizing the overall structure of chromatin, even though it is not part of the histone octamer directly (see Figure 1.1). The presence of histone H1 is believed to be of importance for maintaining a tightly packed structure, and it has been shown that regions enriched in H1 are generally not transcribed, while depleted regions tend to be more accessible and transcribed [35].

lncRNAs There is increasing evidence that long non-coding RNA (lncRNA), non-protein-coding RNA transcripts of more than 200 nucleotides, can regulate the structure and function of chromatin in several ways. One of these mechanisms is via so-called natural antisense transcripts (NATs), which are lncRNAs transcribed from the antisense DNA strand of other RNA transcripts. These complementary sequences can then bind to DNA, via base complementarity or secondary structures, to act as scaffolds to recruit histone-modifying enzymes that themselves lack specific DNA-binding domains. In this way, chromatin modification is thought to be able to act in a site-specific manner [51]. Typically, by recruiting repressive epigenetic factors such as Polycomb complexes, lncRNAs can act in a repressive manner (see also the next section). However, lncRNAs can also be found at independent loci, and be transcribed directly, without antisense transcript mechanisms.

Recently, activating functions of lncRNA have also been found. It has been speculated that a possible mechanism may be through activation in *cis*, by recruitment of transcriptional activators, or by interaction with the Mediator complex to facilitate physical looping between the transcribed lncRNA locus and the target promoter. Indeed, transcription from lncRNA loci has been shown to be highly abundant. Such lncRNA loci have also been shown to be associated with enhancer elements around mammalian genomes, and transcription from enhancers has been suggested to be a general feature [52–54].

X-chromosome inactivation The most studied example of repressive regulation via lncRNAs is X-chromosome inactivation (XCI) in mammals [55]. In XCI, one of the two copies of the X-chromosomes in female differentiating cells become inactivated to compensate for the dosage effect of having twice as many X-chromosome genes as male cells. It was early established that the inactivated X-chromosome is much more compact (Barr body), and that this is the result of a compact chromatin structure via repressive and inactivating chromatin modifications. The inactivating histone modifications are established via the X-inactive specific transcript (XIST), which is a large lncRNA encoded at a region called the X-inactivation center on the X-chromosome of mammals. Multiple copies of XIST are expressed exclusively on the inactive X-chromosome, which in turn binds and coats the X chromosome it is transcribed from. The binding of XIST recruits protein complexes, such as Polycomb, that modify the histones, and remodels the chromatin into a tightly packed heterochromatin structure [35].

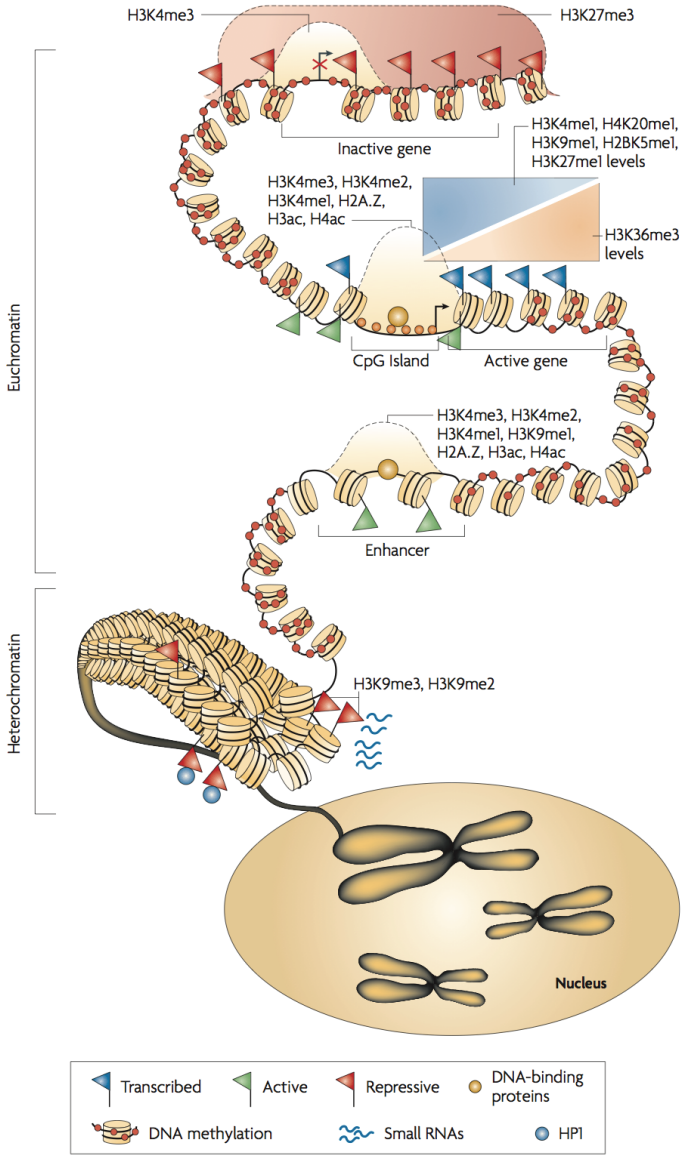

Figure 1.3: Architectural features of chromatin. Broadly, chromatin can be divided into euchromatin and heterochromatin, indicated at the left. Euchromatin constitutes a highly dynamic structure regulated by both activating and repressive histone modifications. Top: Bivalent chromatin, poised for transcription, marked by activating (H3K4me3) and repressive (H3K27me3) histone modifications simultaneously. Middle: Promoter regions are marked by activating histone modifications such as H3K4me1-3 and acetylation, while enhancers are marked by H3K4me1 in addition to other active histone modifications. Actively transcribed genes are marked by H3K36me3 at the 3' end, and by monomethylation of several histones at the 5' end. Bottom: Heterochromatin, consisting of a compacted structure marked by repressive histone modifications such as H3K9me3 and H3K9me2. DNA methylation (red circles) is found throughout chromatin, except at particular CpG islands of active promoters and possibly enhancers. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [38], copyright (2014).

1.1.3 Genomes in 3D

Even though scientists imagined the importance of the three-dimensional (3D) positioning of genomic elements from early on [27], and the FISH technology allowed for visualization of the positioning of specific sites within the nucleus [26], it was not until the introduction of the chromosome conformation capture (3C) techniques [56] that the 3D positioning of genomic elements were fully taken to the post-genomic era. In 3C, and similar high-throughput adaptations, interaction frequencies can be quantified between selected regions in the genome (see section 1.2 for an in-depth discussion).

1.1.3.1 Radial positioning of genomic elements

The emerging evidence of the existence of non-random organization of chromosomes into territories [29], gave rise to the speculation that the relative positioning of chromosomes and genes was important for how the genes were regulated. One of the earliest indications that this could be the case came from studying the relative positions of human chromosomes 18 and 19 using FISH. Chromosome 18 is known to be gene-poor, low in CpG content, and generally associated with low-activity histone marks, while chromosome 19 on the other hand is known to be generally active and gene-rich. Interestingly, chromosome 19 was shown to occupy central parts of the nuclear space, while chromosome 18 was shown to be positioned towards the periphery [57]. The tendency of gene-dense chromosomes to be more central, while gene-poor chromosomes are more peripheral has later been confirmed for all human chromosomes [58]. Additionally, the same pattern has been seen for gene-dense and gene-poor parts of chromosomes [59]. This radial positional principle was later shown for individual loci, such as α-globin or HoxB, as well [60, 61], where genes are repositioned towards the centre of the nucleus upon activation. The emerging picture is much more complex, however, since many examples of gene repositioning without clear regulatory effects have also been found [62].

1.1.3.2 Transcription factories

Another important aspect related to radial gene positioning is the idea that genes can cluster themselves spatially within the nucleus in order to be transcribed in a concerted fashion (see Figure 1.4). This type of gene arrangement, called a transcription factory, has been speculated to be formed as a way for genes to reposition themselves in local regions with high concentration of polymerases [63–66]. The existence of transcription factories today is not disputed, however the exact consequences of these gene clusters are not completely understood [67]. Current evidence suggests that somewhere between a few hundred to several thousand transcription factories form for each cell, with a clear enrichment of elongating polymerases at each site [67, 68]. The factories are thought to form during cell-differentiation upon activation, and remain even after the genes in the factory are no longer active [69]. Genes found to be proximal have been shown to be more correlated in terms of their expression [70], and simulation studies have shown that it is theoretically possible for as many as 80% of co-expressed genes to be spatially proximal [71]. Evidence of specialization of individual transcription factories, such as clustering of genes that belong to the same pathway has been sparse [67]. However, spatial clustering of active globin genes in mouse and human cells has

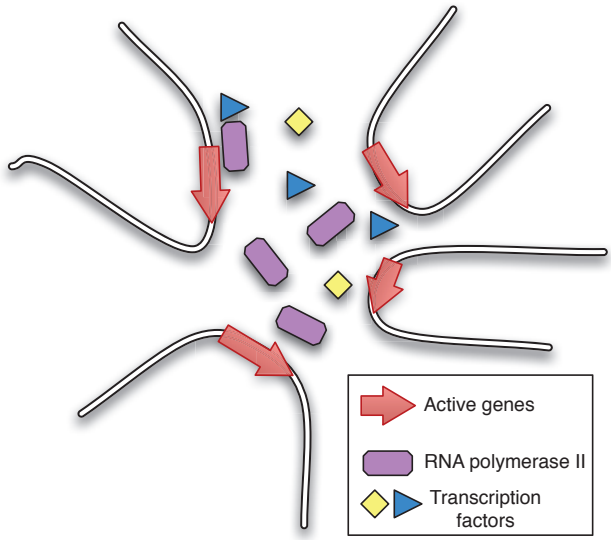

Figure 1.4: Transcription factory. Clusters of active genes (red arrows) found proximal in the genome, sharing access to RNA polymerase (purple) and transcription factors (blue and yellow).

been reported [66].

Evidence of specialization of transcription factories was also found in a study where mouse globin genes in erythroid tissues were analyzed using a modification of the chromosome conformation capture-technique, utilizing microarrays for quantification (see section 1.2.4) [72]. In that study, the authors found evidence for clustering and co-localization of genes regulated by the erythroid transcription factor Klf1. Interestingly, by using knock-out mice, the authors were also able to show that Klf1 was required for the formation of Klf1-associated transcription factories.

A well-known transcription factory-like organization of transcribed ribosomal RNA (rRNA) genes, localized in the nucleolus, has been studied extensively. The nucleolus is a dense structure found within the nucleus of eukaryotic cells, which forms at specific chromosomal regions. In the early 1960s, it was shown that the nucleolus is the site of ribosomal RNA synthesis and processing, in addition to ribosome production [73]. In human cells, this happens via clustering of $\sim$40 rRNA genes found on five different chromosomal locations that are transcribed in a concerted fashion by RNA polymerase I [74].

Interestingly, recent studies have characterized nucleoli as more than simply rRNA synthesis loci. In a study where DNA associated with nucleoli was sequenced genome-wide, the authors identified 97 chromosomal regions, encompassing around 4% of the human genome, as associated with nucleoli. These regions, called nucleolus-associated chromatin domains (NADs), were shown to be associated with repressive histone marks and lower gene expression. Some of the regions showed overlap with regions previously shown to interact with the repressive environment of the peripheral part of nucleus, called the nuclear lamina (discussed next), indicating that repressed genomic regions categorize into several distinct classes [48, 75, 76].

1.1.3.3 Nuclear lamina interactions

The nuclear lamina has long been recognized as important for the structure of the chromatin fibre [77, 78]. The nuclear lamina is a protein meshwork associated with the inner membrane of the nucleus, particularly rich in filament proteins called A- and B-type lamins defined according to sequence homology and biochemical properties. The lamins interact with proteins bound to the nuclear membrane and provide both structural and regulatory functions. A-type lamins are expressed in a cell-type specific manner, while B-type lamins seem to be more or less constitutively expressed. Additionally, it has been shown that A-type lamins are present throughout the nucleus, while B-type lamins seem to have a more stable localization at the nuclear periphery [79, 80]. Experiments where genes have been artificially tethered to the periphery have been conducted, but with varying results, since de-activation of genes was sometimes seen [81, 82], but other times not [83].

Consistent with the finding that inactive, gene-poor parts of chromosomes tended to reside towards the nuclear periphery, Guelen et al. showed that maps of B-lamin-associated regions throughout the genome were consistently gene-poor, had low gene expression, and were associated with repressive histone modifications [84]. Interestingly, these authors also showed that the lamina-associated regions occurred in domains of typical size in the range 0.1-10 megabases. The authors called these regions lamina associated domains (LADs). They also found that the borders of the LADs were demarcated by CTCF, suggesting that LAD structure could be partially defined by binding of insulator elements.

1.1.3.4 Domain organization of the genome

The idea that chromosomes are organized in a domain architecture has been considered ever since the discovery of differential staining patterns and their relationship with the underlying GC-content. Originally, such segmentation was attributed to the theory of "isochores", where local differences in stretches of alternating GC-content along mammalian chromosomes were described. It was early noted that gene-density tended to be higher at regions with high GC-content, and that such gene dense regions were occupied by more actively expressed genes [48].

Compartments The domain organization of the human genome was confirmed and expanded in a landmark paper by Lieberman-Aiden and colleagues in 2009 when they were able to probe the structure of the entire human genome by coupling chromosome conformation capture to high-throughput sequencing [85], using a method called Hi-C (discussed in section 1.2.6). In Hi-C, contact frequencies between bins covering the entire genome are quantified using paired-end sequencing, giving interaction matrices of all-versus-all bins. One of the most striking results in this study was that the human genome appeared to be organized into two separate compartments that the authors called A (or open) and B (closed) compartments (see Figure 1.5). These compartments were larger than the previously described LADs, with a size range between $\sim$1-10 megabases. Intriguingly, the authors demonstrated that genomic interactions formed mostly within compartments, and much less frequently between compartments. Additionally, the A compartments were associated with open chromatin and gene-dense regions, while B compartments were associated with gene-poor, inactive regions.

In a follow-up study, re-analyzing the same Hi-C data with improved computational methods, the authors revealed that in addition to a binary categorization into A and B compartments, patterns of interaction frequencies was found related to underlying properties of chromosomes in a more continuous manner. For example, the relative positions along chromosome arms, with increased interaction frequency near centromeres and near telomeres, were found to be important factors, confirming earlier studies showing that centromeres (as well as telomeres) tended to be co-localized in 3D [86].

In a similar study, Yaffe et al. [87] showed that human chromosomal regions could be categorized into three classes of interaction frequencies, one defined by high activity and corresponding well with previously defined active domains, while inactive chromatin could be subdivided into two states defined according to the relative positioning on chromosome arms. One of the inactive classes was found to be close to centromeres, while one was close to telomeres.

TADs Recently, using Hi-C at a much higher resolution than the original study, Dixon et al. where able to show that the genome is folded into further domain structure, called topologically associating domains (TADs), within the previously discovered compartments [88]. These $\sim$900kb sized domains were found to occupy $\sim$91% of the genome, and are characterized by a much higher within-domain interaction frequency compared to between domains (see Figure 1.5). Interestingly, the authors found that the domain borders were demarcated by CTCF, similar to the previously described LADs. The TADs were found to be related to, but not identical to the LADs. For example, the authors found that the TAD boundaries often marked the transition between LAD and non-LAD parts of the genome, or A and B compartments. Mapping of TADs was performed for undifferentiated and differentiated cells from both human and mouse, to be able to compare across the cell-lines. The TAD organization proved remarkably stable across the cell types, and was also highly conserved between human and mouse, suggesting an evolutionary conserved and important function of this organization.

Even further, lower-order domain organization, into so-called sub-TADs, has been described [89]. Phillips-Cremins and colleagues applied a high-resolution chromosome conformation capture technique, called 5C (discussed in section 1.2.5), to zoom in on the structures within larger TADs at selected regions in the genome. The identified sub-TADs were found to sometimes be cell-type specific, and other times constitutive. The authors proposed that binding of CTCF and cohesin results in the creation of boundary elements between TADs, but additionally that CTCF and cohesin create subdomains (sub-TADs) by anchoring constitutive interactions around genes with cell type-specific expression.

In a recent study, Nora et al. [90] studied and compared TAD composition around the X-inactivation center (see section 1.1.2.3) in murine cells. In that study, the authors found that disruption of a boundary between two TADs caused both changes in contacts around this boundary, and misregulation of associated genes. Interestingly, during cell differentiation, the authors noticed that some individual TADs became associated with the lamina (LADs, see section 1.1.3.3). This led the authors to propose that TADs constitute a modular framework where chromatin structural changes can occur. Also, gene-expression profiles were found to be highly correlated within TADs, as opposed to between TADs, indicating that regulation of genes via cis-regulatory elements may happen in a coordinated fashion within TADs.

Taken together, these findings hint at the importance of such genomic domain organization in shaping the local regulatory landscape of genomes.

Replication timing and replication domains The replication program of eukaryotic chromosomes is a tightly regulated process where multiple replication origins initiate replication at specific times in an orderly progression throughout the S phase of the cell-cycle. The relationship between GC-rich regions, transcriptional activity and replication timing has long been known [92], and it was early noted that late replicating regions tend to be positioned near the periphery of the nucleus, while early replicating regions is more randomly localized [93]. Also, the fact that the replication program is organized in domains has been recognized as an important factor [94].

With genome-wide high-throughput genomic methods, replication timing profiles have been mapped extensively and comprehensively throughout many tissues and across species. In such studies, the domain organization of replication timing has been confirmed. Interestingly, by comparing across several tissues, Hansen et al. [95] were able to distinguish domains that were constant across cell-lines, and domains that were dynamic ("plastic"). Remarkably, the plastic domains were found to cover almost 50% of the human genome, hinting at a tightly regulated cell-type specific replication timing for a large part of the genome. Further analysis of genome wide replication timing profiles has revealed that constitutive replication timing profiles are evolutionary conserved across human and mouse [96]. Interestingly, however, also cell-type specific replication timing changes were found to be similar when comparing related cell types across species. These cell-type specific changes were, in both species, found in units of 400-800 kb.

Ryba et al. also compared the replication timing profiles with Hi-C data from Lieberman-Aiden et al. [85], and found a remarkable correlation with the A and B compartments, including cell-type specific patterns. The correlation between Hi-C and replication timing was the strongest signal identified of all epigenomic features compared, leading them to conclude that replication timing occurs in spatially separate nuclear compartments. Similar results were later reported by Yaffe et al. [97]. Several authors have also found a striking 3D co-localization between DNA replication origins, again indicating the tight relationship between chromatin 3D structure and replication [98, 99].

Domains as a basis of genome regulation The correlations of domain structure across cell-types and species from various sources of epigenomic data has led to the need of a unifying view of chromatin organization. Recently, several authors have posed TADs as the candidate for the structural basis for large parts of the regulatory landscape of genomes [100]. The current view states that groups of active TADs combine to form A compartments, while groups of inactive TADs combine into B compartments (see Figure 1.5).

Certain histone modifications (notably repressive modifications such as H3K9me2 and H3K27me3) tend to be found in blocks of similar sizes as TADs [101, 102]. The fact that boundaries of TADs and repressive histone modification domains often coincide hints to a mechanistic link between the two. Also, borders of LADs often coincide with borders of TADs [88], and disassociation of entire TADs from the nuclear lamina could explain why differences in LAD structure between cell-types often are found in TAD-sized units. A similar

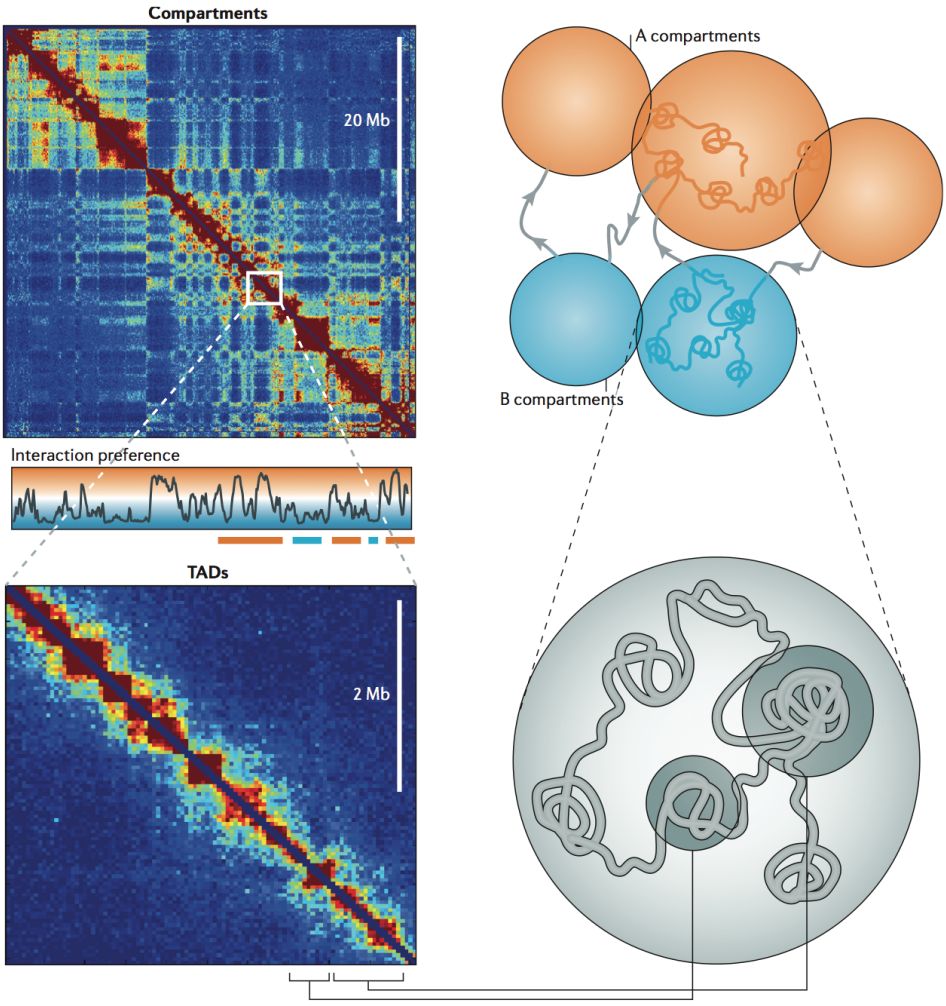

Figure 1.5: Domain architecture of the genome. Top left: Heat map visualization of a Hi-C dataset from a selected genomic region, with compartments illustrated by the interaction preference between regions in the genome. Top right: Schematic illustration of A and B compartments, showing high degree of interactions between A compartments, and similarly between B compartments, but few interactions between A and B compartments. Bottom left: Zooming in on one of the compartments allowing visualization of TADs recognized as squares along the diagonal of the Hi-C heat map. Bottom right: Schematic illustration of the chromatin topology of TADs, showing high degree of interactions within TADs, but few interactions between TADs, resulting in smaller domains within each compartment. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [91], copyright (2014).

mechanism could explain why differences in replication timing happen in similarly sized units [100].

1.1.3.5 Boundary elements and genome organization

CTCF CCCTC-binding factor (CTCF) is a constitutively expressed protein containing an 11-zink finger DNA-binding domain. Using high-throughput methods, binding sites of CTCF have been mapped extensively in several tissues and species. Such studies have revealed up to ∼50000 binding sites throughout the genome, approximately half of which seem to be shared between cell types [103–105]. While CTCF initially was considered mostly as a transcription factor capable of activation or repression of gene expression, today a large number of important regulatory functions of this protein has been identified.

One of the most prominent features of CTCF is its ability to function as an insulator element by blocking physical interactions between enhancers and promoters, and to prevent the spread of repressive chromatin (H3K27me3) into surrounding genomic regions [106]. This insulator role of CTCF was confirmed when considering genome-wide binding of CTCF, even though only a small subset of bound CTCFs was found at repressive chromatin boundaries [105].

The enhancer-blocking activity of CTCF has also been studied in light of genome-wide CTCF binding data. In a computational study, Xie et al. compared the correlation of gene expression of pairs of genes separated by CTCF sites with pairs of genes not separated by CTCF sites. In the study, the authors found that gene pairs not separated by CTCF were much more correlated than CTCF-separated genes, which had correlation close to background level [107].

Recently, a third role of CTCF has been emerging. Several studies comparing genome-wide 3D interactions and CTCF-binding sites have shown that sites bound by CTCF are co-localized in 3D, suggesting that CTCF can act as a facilitator of chromatin contacts [108, 109]. In a pioneering study, Handoko et al. used a technique called ChIA-PET (discussed in section 1.2.7) to map physically interacting CTCFs genome wide in mouse embryonic stem cells [110]. They identified ∼1500 intrachromosomal and ∼300 interchromosomal interactions between sites bound by CTCF across the genome. Interestingly, they revealed that epigenomic patterns at the boundaries of interacting pairs of CTCFs were markedly different from non-interacting CTCF pairs. By comparing the identified CTCF-loops with LADs, they identified a depletion of loops within LAD units. At LAD borders on the other hand, loops seemed much more enriched, indicating that CTCF binding is associated with LAD formation.

In addition to being involved in facilitating 3D contacts, CTCF has been shown to be capable of regulating transcriptional pausing of polymerase at specific promoters. Such activity is thought to be caused by CTCF's ability to stabilize long-range interactions that can interfere with the elongation of the polymerase [40].

Binding of CTCF has been shown to be regulated both by epigenetic and protein binding interaction mechanisms. Specifically, DNA methylation of the regions around CTCF-binding sites has been shown to repress binding of CTCF, thereby suggesting a mechanism for cell-type specific control of the action of CTCF. Interactions between CTCF and other regulatory

proteins may be yet another mechanism by which CTCF-activity can be controlled [40].

Cohesin Several papers in 2008 reported that CTCF binding sites were co-occupied by cohesin, a protein complex known to regulate the cohesion of sister chromatids during cell division [111–113]. This was a major indication that CTCF and cohesin together were responsible for establishing long-range 3D contacts in the genome, and the characteristic ring-like structure of some of the subunits of cohesin even indicated a possible function of cohesin as a stabilizer of chromatin loops via trapping the two DNA strands inside the ring [114]. It was observed that cohesin could bind to DNA even after CTCF had been depleted, but with reduced affinity and with non-specific positioning, indicating that CTCF acts as a recruiter of cohesin which in turn is responsible for chromatin loop formation and insulator functionality [113].

This theory was refined as Zuin et al. interrogated the effect on physical chromatin interactions by depleting cells of either cohesin or CTCF [115]. In that study, a general loss of chromatin 3D interactions was seen upon depletion of cohesin, but without affecting the borders of TADs. Interestingly, however, depletion of CTCF both reduced the occurrence of 3D interactions and increased the interactions between TADs. This indicates that cohesin has a main role in establishing 3D interactions within TADs, while CTCF is important for the segregation of TADs. CTCF and cohesin binding cannot be the only factors establishing boundaries of TADs, however, since only $\sim$15% of CTCF binding sites across the genome are associated with TAD boundaries, and the rest are present within TADs. It has been suggested that CTCF binding within TADs is primarily involved with mediation of cell-type specific short-range contacts [40].

1.1.3.6 Cis-regulatory interactions

In addition to the insulator-mediated loops responsible for formation of domains that were discussed in section 1.1.3.5, at least three other classes of physical interactions in cis occur in mammalian genomes [116]. These additional three classes are responsible for transcriptional activation, transcriptional repression and recycling of transcription, respectively, and will be discussed separately in this section.

Enhancer-promoter interactions - transcriptional activation One of the most widely studied classes of genomic 3D interactions has been the interactions between regulatory elements, such as promoters and enhancers, causing activation of transcription in tissues where interactions occur. The exact mechanism of how promoters and enhancers could combine to activate transcription was debated at the end of the 20th century [117, 118]. In the post-genomic era, and with the arrival of the 3C-based techniques, however, the looping model seems to be favored [119]. In this model, proteins, including transcription factors with affinity to motifs on the DNA, bind to the enhancer region, forming an enhanceosome which has affinity to proteins bound at the promoter. This causes them to form a loop between themselves via binding of the two protein complexes. Additionally, the activation of transcription at the promoter often requires the binding of coactivator proteins (without having DNA-binding competency themselves) to the promoter site to act as histone modifiers to further enhance the

activity of the promoter [44]. Also, enhancers can themselves recruit the basal transcriptional complex (including polymerase II), poising the activation of the promoter target until activation signals cause the complex to be transferred to the promoter. The action of enhancers can occur at distances as large as 100 kilobases or more, and can occur either upstream or downstream relative to the promoter, and can even be positioned within the transcription unit itself. Enhancers are also able to activate multiple promoters, and can combine with other enhancers to activate a single promoter [35]. Interestingly, recent evidence based on 5C methodology (discussed in section 1.2.5), applied to 1% of the human genome, has shown that only $\sim$7% of looping interactions are with the nearest gene [120]. Additionally, the same article noted that even though enhancers could be located in any direction relative to the promoter, a bias towards elements being located $\sim$120 kilobases upstream of the promoters were found. The same article also found evidence for several complex networks of interacting promoters and enhancer elements with functional effects on gene expression.

The observation that looping between active genes and regulatory elements often involves several genes and regulatory elements led to the "active chromatin hub" (ACH) theory. In this model, spatial units of multiple regulatory DNA elements, together with corresponding genes, cluster at certain sites, in effect establishing independent expression domains [121].

One of the most well-studied systems of the effect of cis-interactions is the mouse β-globin locus. At this locus, several studies where 3C (discussed in section 1.2.3) has been applied show interactions between an upstream regulatory region called the Locus Control Region (LCR) and active β-globin genes, while no such interactions have been found in tissues where these genes are not expressed [122]. Interestingly, the LCR contains several DNase I hypersensitive sites (HS) that modulate different expression patterns of the four different β-globin genes throughout development [123], all controlled via looping of different HSs to different genes at different developmental stages [124]. Similar LCR elements have been found in a large number of other regions, including the α-globin cluster, the major histocompatibility locus, the immunoglobulin heavy chain locus, and many others [125]. Another notable example is the T_H2 LCR, where interactions between a promoter region of the IFN-γ gene on chromosome 10 is controlled by regulatory regions on a separate chromosome [126]. While such interchromosomal regulatory interactions are rare, several examples have been found [127–129].

Several genome-wide studies of promoter-enhancer interaction, utilizing 3C technology coupled with next-generation sequencing (see sections 1.2.6 and 1.2.7), have been performed. In a recent study, a high-resolution dataset of interactions in a human fibroblast cell-line (IMR90) revealed that transient enhancer-activation in the same cell-line did not induce looping between those enhancers and their promoter targets. Instead, loops between enhancers and promoters seemed to be present before activation. By comparing between cell-type specific enhancers in human embryonic stem cells (hESC) and IMR90, however, promoter-enhancer interactions seemed to be highly cell-type specific. This led the authors to speculate that cell-type specific promoter-enhancer looping forms an additional layer of regulation determining the actual transcriptional outcomes in the different cell-types.

Regulatory interactions do not necessarily occur between a single enhancer and a promoter, however. For example, Li et al. [130] used ChIA-PET to map regulatory interactions involving RNA polymerase II (RNAP II), and found extensive and widespread clusters of

promoter-centered chromatin interactions throughout the genome for several different cell-lines. In that study, most interactions were found to be involved in a complex of interactions between several promoters and enhancers.

Polycomb-mediated looping - transcriptional repression As was discussed in section 1.1.2.3, the Polycomb complex induces the formation of closed and repressive chromatin by promoting the spread of repressive histone modifications. Interestingly, examples exist where repression of Polycomb is mediated by complex higher-order structures, where Polycomb response elements (PRE) and repressed genes are clustered to impose a repressed state [131]. Using 4C (discussed in section 1.2.4) Bantignies et al. [132] demonstrated that two Hox loci in *Drosophila* separated by 10 megabases on the same chromosome were co-repressed by association with PcG proteins. This repression was shown to be caused by looping of these two sites mediated by PcG, contributing to the specification of body structures in *Drosophila*. Similar repressive associations of PcG and distal repressed regions have also been found in human cells [133, 134]. The purpose of forming such PcG chromatin hubs has been speculated as serving to create nuclear compartments depleted of transcription factors and RNAP II in order to bring about stable maintenance of chromatin silencing [132].

Promoter-terminator looping A third, but much less understood class of chromatin looping interactions, are interactions between the promoter and the 3' terminator sites of genes transcribed by RNAP II. Initially, such loops were described in yeast [135], and were speculated to provide an efficient way for the same polymerase molecule to reinitiate at the promoter site right after transcriptional termination [136]. Such polymerase re-cycling has been speculated to be essential for sustaining continued transcription for certain loci [137].

Similar interactions have been described in mammalian cells. For example, the well-studied tumor supressor gene BRCA1 has been shown to form loops between the promoter and terminator site [138]. In mammals, however, the mechanisms and regulatory consequences of promoter-terminator loops seem much more complex [119]. Even though it has been speculated that similar re-cycling mechanisms are present in mammals, other theories such as maintenance of repressed states [138], regulation of elongation and splicing [139] and maintenance of active transcription, have been proposed [140].

1.1.3.7 The dynamic genome

A fundamental property of the three-dimensional architecture of the genome, is its dynamicity and variability across cells, caused by differences in cell-cycle progression, differentiation stage, transcriptional status and general stochasticity [141–143].

While large-scale chromatin motion is usually observed to be partially constrained [144], FISH analyses of selected loci typically show a high degree of variability of genome organization across cells [116, 145, 146].

With 3C-based technologies, however, the resulting data are derived from an average of millions of cells [146]. Computational modelling of chromatin structure based on such data can be used to quantify the degree of variability of structures, and typically shows that clusters of structural ensembles are needed to explain the observed averages [146, 147]. Techniques

such as TCC (discussed in section 1.2.6), where technical noise in the data is reduced, still show that a population of structures is needed to explain the data [148].

Recently, a novel technique called single cell Hi-C (discussed in section 1.2.6), made it possible to map genome-wide 3D interactions in individual cells [149]. In that study, Nagano et al. performed single cell Hi-C on 60 different mouse CD4+ cells, and found that TADs were consistently conserved across the cells. Importantly, the TADs mapped in single cells were shown to correspond to TADs found using conventional "ensemble" Hi-C methodology. However, even though interactions within TADs were highly conserved, interactions between TADs (inter-domain interactions) were found to be variable between the cells. Also, the results indicated that each chromosome contacts a limited and constant number of other chromosomes in single cells, but with high variability across cells.

The dynamicity of chromosomes does not only manifest itself across different cells, but also over time during the cell-cycle. While the structure of the genome during interphase has been found to be highly compartmentalized into for example TADs, the structure of chromosomes during mitosis has until now only been studied in the microscope. However, recently, Naumova et al. [150] mapped the structure of human chromosomes in different cell-cycle stages of HeLa cells using both Hi-C and 5C. In that study, two distinct folding states were found. While domain-type architecture, such as A and B compartments and TADs, were present during interphase as had previously been shown, the metaphase chromosomes adopt a folding state where the domain architecture is completely lost. To account for the observed metaphase structures, which where shown to be similar across all chromosomes, the authors proposed a chromatin architecture where the chromatin fiber is linearly organized and compacted in a two stage process.

1.1.3.8 Chromatin and disease

Mutational events at enhancer elements provide one of the most direct links between chromatin architecture and outcome of disease. Several examples of mutations at enhancer elements found distal from the affected genes have been found. For example, a single nucleotide polymorphism (SNP) at an enhancer element found $\sim$335 kb from the MYC proto-oncogene target has been shown to increase binding of a transcription factor at the enhancer element. This increased binding affinity was shown to enhance expression of the MYC gene in colorectal cancer [151]. Other mutational events, such as insertions/deletions and structural variations, have also been shown to alter gene expression by affecting the regulatory targets of genes [152].

Several links between the three-dimensional architecture of chromatin and disease have been made in the recent years. For example, by using Hi-C data, Engreitz et al. were able to show that regions involved in translocation events were proximal inside the nucleus, pointing to a causal role of chromatin contacts and translocations [153]. For example, specific translocation partners often found in certain cancers were found to be closer to each other than expected by chance. In that study, the authors found evidence for both tissue specific and constitutive features of chromatin structure determining rearrangements in human disease.

Similarly, it was shown in two independent studies that somatic copy-number alterations in cancer are highly correlated with spatial proximity [154, 155]. In both papers, the authors

showed that genomic regions found spatially proximal in the nucleus were more likely to form copy number alterations. Since copy-number alterations are one of the most common alterations found in cancer, this points to a central role of genomic 3D architecture in cancer formation and outcome.

Additionally, somatic mutation events have been shown to be highly correlated with chromatin features and organization. Specifically, it was recently shown that compact, heterochromatin-associated parts of the genome was much more likely to exhibit somatic mutations across several cancer cell-lines [156]. Again, this could point to causal links between chromatin structure and mutational events.

Large-scale chromatin architecture alterations involving changes in heterochromatin have also been studied. Particularly, mutations in genes coding for lamina proteins (see section 1.1.3.3) have been linked to several diseases known as laminopathies. In these types of disorders, a range of clinical symptoms have been described, including muscular, neurological, lipodystrophic and related to accelerated ageing [157]. Several of these disorders, caused by mutations in one of the lamina proteins, have been shown to significantly alter heterochromatin composition by causing abnormalities in the nuclear periphery [158].

Several additional examples of epigenetically linked neurodevelopmental and neurodegenerative disorders have also been mapped in the recent years [159]. It has been speculated that such disorders arise due to misregulation of chromatin composition and structure in the brain [160].

1.2 Molecular techniques

In the last decades, a range of novel techniques for analyzing the architecture of chromatin has been developed. Particularly, the coupling of next-generation sequencing to traditional chromosome conformation capture techniques has revolutionized the possibilities of exploring genomic 3D interactions in detail [161]. In this section, the most important methods for studying genomic 3D architecture will be discussed, focusing on the 3C-based technologies that are most relevant to this thesis.

1.2.1 Fluorescence in situ hybridization (FISH)

FISH is a combined molecular and cytological approach where fluorescently labeled DNA probes are hybridized to complementary sequences on chromosomal preparations fixed on slides. The probes are then visualized using microscopy. FISH has been available for several decades, but is still widely used both in research and diagnostics. The wide usage of FISH is attributed to the fact that it provides spatial information at intermediate degree resolutions in single cells, and that it is relatively easy to apply. Techniques utilizing FISH are still being refined and diversified into more specialized versions, and the three most widely used FISH variants in 3D genome analysis are Cryo-FISH, 3D-FISH and Immuno-FISH, which are summarized below. This section is based on descriptions given in [162].

1.2.1.1 Cryo-FISH

In Cryo-FISH, ultra thin cryosections of fixed cells are analyzed using two-dimensional microscopy. This allows for much better resolution than conventional FISH methods, and has therefore been used for validation of findings based on long-range interactions found using 3C-technologies. Cryo-FISH can visualize spatial relationships between chromosome territories.

1.2.1.2 3D-FISH

3D-FISH uses cross-linking to preserve nuclear architecture, and visualization of results requires confocal laser-scanning microscopy. 3D-FISH can visualize the relative positioning of chromosomes and sub-chromosomal regions, and has been used to visualize all human chromosomes inside nuclei [163].

1.2.1.3 Immuno-FISH

In Immuno-FISH, standard FISH is combined with immunofluorescence to visualize both DNA and proteins simultaneously. This method can be used to investigate co-localization of genomic elements in subnuclear bodies in interphase nuclei.

1.2.2 Next-generation sequencing

With the advent of next-generation sequencing, traditional Sanger-based sequencing techniques, based on radioactively (or fluorescently) labeled dideoxynucleotides, have in many cases been replaced. Next-generation sequencing approaches allow the sequencing to be done in parallel such that millions of sequences can be interrogated simultaneously based on clonal amplification of DNA fragments. To achieve this, the sequences are often spatially separated on plates or slides, and interrogated using a high-resolution camera. Several technologies are available, including 454 (Roche Diagnostics), SOLiD (Applied Biosystems), and Solexa (Illumina). Most next-generation sequencing technologies allow for paired-end sequencing, where the two ends of the same DNA molecules are sequenced from both sides [164].

1.2.3 Chromosome conformation capture (3C)

In the following sections, the technologies based on chromosome conformation capture (3C) will be discussed. The underlying concept of all these methods is the quantification of ligation junctions by digestion and re-ligation of fixed chromatin in cells, and that the quantified DNA contact frequencies reflect proximity within the nucleus [165]. 3C was invented in the early 2000s by Dekker et al. [56], and originally, contact frequencies were quantified using quantitative PCR (qPCR), but today paired-end sequencing is used to a much larger extent. In contrast to the microscopy-based FISH techniques mentioned above, 3C-based methods allow for much more systematic and quantitative characterization of genome topology, at much higher resolution than FISH. The drawback, on the other hand, is that 3C-based methods are mostly performed on large populations of cells, thereby losing the information at the single-cell level [165].

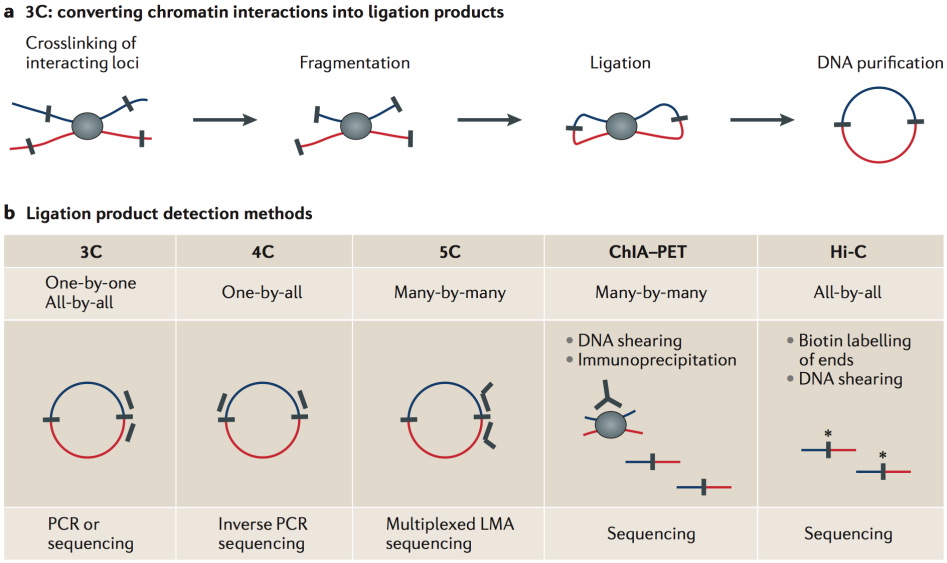

Figure 1.6: (a) Initial steps in the chromosome conformation capture (3C) procedure. (b) Schematic illustration of key concepts for the various 3C-based technologies. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics [91], copyright (2014).

All 3C-based techniques start with the same steps, where the goal is to isolate DNA fragments in spatial proximity. The first step is fixation (cross-linking) of chromatin, often by using formaldehyde. This causes chromatin segments in close spatial proximity to covalently link to one another. The fixed chromatin is then cut with a restriction enzyme (HindIII, SacI, EcoRI, DpnII and others), chosen such that the frequency of cuts provides the resolution necessary for the given analysis. The sticky ends of the fragmented cross-linked DNA are then re-ligated under diluted conditions to favor intramolecular ligation of the cross-linked fragments. The re-ligated DNA molecules thereby form a hybrid consisting of two DNA fragments from the two segments that were cross-linked. After DNA purification, qPCR or sequencing can be used to quantify the number of such hybrid DNA-molecules.

In 3C, primers are designed near the ends of the restriction fragments of interest, enabling quantification of selected ligation junctions. Ligation frequencies, as measured by the amount of ligation-product between the selected primer combinations, are then used to infer spatially proximal fragments [91, 165]. 3C therefore allows for focused quantification of contact frequencies at selected regions in a one-versus-one fashion (see Figure 1.6).

It is important to recognize that DNA sequences close to each other in linear genomic space will by necessity be much more likely to cross-link and form ligations. Therefore, the quantification step also involves determining if DNA segments contact each other more than expected simply due to the (linear) genomic proximity between them. In addition, over large genomic distances, ligation products become very infrequent, and quantification using qPCR becomes infeasible. Also, the differences in primer efficiencies need to be controlled for, by making a control template with all ligation products in equal amounts [166].

1.2.4 Chromosome conformation capture-on-chip (4C)

As the name suggests, chromosome conformation capture-on-chip (4C) was originally applied using microarrays in the quantification step. In 4C, contact frequencies from a given genomic site ("viewpoint") towards all other sites are quantified, thereby making it a one-versus-all type of analysis, in contrast to the one-versus-one nature of 3C.

In 4C, all the regions cross-linked with the viewpoint are amplified via inverse PCR by using two bait-specific primers oriented towards the outer restriction sites. This allows for quantification of all the DNA fragments that were spatially proximal to the bait sequence. The abbreviation 4C is actually used for two similar techniques that only differ in the steps after de-cross-linking. In chromosome conformation capture-on-chip, the ligation junctions are cut using a frequently cutting secondary restriction enzyme prior to re-ligation, such that self-ligated circular structures are created. Inverse PCR is then used to amplify the outer ends of the captured DNA fragments. In a similar technique, called circular chromosome conformation capture (also abbreviated 4C), the formation of circles happens during the 3C ligation step, requiring that both ends of the bait fragments ligate to both ends of the captured restriction fragments. For both methods, quantification after inverse PCR is performed either with microarray analysis or by using sequencing techniques. More recently, the quantification is performed using next-generation sequencing, and is often referred to as 4C-seq [167]. The resulting data consist of a genome-wide profile of ligation events with peaks corresponding to significant interactions. Again, the unspecific background signal needs to be filtered out, often by using replicate libraries and by comparison to the signal of relative abundance in a local area around each peak [168].

1.2.5 Chromosome conformation capture carbon copy (5C)

In chromosome conformation capture carbon copy (5C), the goal is to capture all interactions between a set of selected regions, in a many-versus-many fashion. 5C requires specifically designing primers (5C primers) that anneal to the ends of the restriction fragments. A ligation between two restriction fragments in the 3C library will therefore result in two 5C primers annealing adjacent to each other on each side of the ligated restriction sites. Annealing and ligation of the primers is performed in a multiplexed fashion by using thousands of 5C primers simultaneously. Amplification via universal primers attached at the ends of all 5C primers is then performed. Finally, the ligation products can be quantified using microarrays or next-generation sequencing. The resulting data will be a matrix of interaction frequencies between the two sets of restriction fragments selected prior to analysis. Again, also for 5C data, the fragments that are close in linear genomic space will have a higher chance of forming contacts, often revealed as a diagonal along the resulting data matrix. The method and its resolution is restricted by the possibility of designing appropriate primers at the ends of the restriction fragments. In practice, the resolution is also restricted by the need to use many primers simultaneously, and is therefore suited mostly for focused study of selected regions [91, 165].

1.2.6 Hi-C

Hi-C was introduced by Lieberman-Aiden et al., and constituted a major breakthrough in the study of chromatin architecture, since it allowed for genome-wide quantification of interactions in an all-versus-all fashion [85]. In Hi-C, an extra step is needed in the standard 3C-protocol after the restriction enzyme cutting has been performed. In this extra step, the sticky ends are filled in with biotin-labeled nucleotides, prior to purification and shearing of the DNA. The biotin marks are then subsequently used to selectively purify the hybrid DNA-molecules, by pulling down on the biotin mark. The resulting library of hybrid DNA molecules, consisting of one fragment from each of the two ligated regions, is then sequenced using paired-end sequencing (see Figure 1.6). By subsequently mapping each end-pair back to the reference genome, a genome-wide aggregated contact matrix is obtained. Due to the extremely high throughput needed to fill a matrix of all-versus-all restriction fragments, the interaction frequencies are usually aggregated over equally-sized bins of a certain size. At the moment, the human genome can routinely be mapped at resolutions of 100 kilobases [85]. However, for focused analysis of shorter intrachromosomal interactions, resolutions as low as 40 kilobases have been reached [88]. Various biases influence the row- and column sums in the resulting matrix, and correcting and normalizing these data is necessary (see section 1.3.1). A similar genome-wide technique, called genome conformation capture (GCC), has been applied for mapping of yeast chromosome interactions as well [169].

A number of related techniques similar to Hi-C have been developed. One technique, called tethered chromosome conformation capture (TCC), tethers the cross-linked structures on the surface of straptavidin-coated beads. While in this solid phase, sticky ends are marked with biotin and ligated. The cross-linking is then reversed, and DNA is purified and hybrid fragments are pulled down based on the biotin mark. By cross-linking DNA fragments during the ligation, the signal-to-noise ratio is enhanced [148].

Recently, a single-cell version of Hi-C was published [149]. In single cell Hi-C, cross-linking, restriction enzyme digestion, biotin fill-in and ligation is all done inside individual nuclei, in contrast to traditional (ensemble) Hi-C, where these steps are performed after cell lysis and dilution. Then, by isolating individual nuclei, DNA-ligation events can be quantified for each nucleus separately. Since this method is done on individual cells, theoretically a maximum of two ligation events per restriction fragment is possible (on autosomes). There will therefore be a theoretical maximum number of ligation events that are possible to quantify for a given restriction fragment [170].

1.2.7 ChIA-PET

In chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) [171], the goal is to identify all interactions between regions bound by a protein of choice, effectively a many-versus-many type of analysis (see Figure 1.6). In ChIA-PET, after cross-linking, sonication is used instead of restriction enzyme digestion, followed by chromatin immunoprecipitation (ChIP) to selectively pull down on a specific DNA- or chromatin-binding protein using an antibody. Biotinylated linkers are then added to the fragment ends containing restriction sites specific for Mme1, which are then ligated (under dilute conditions). The cross-links are then reversed, and the Mme1 restriction enzyme is used for digestion, and hybrid-

fragments (containing biotin) are then captured and quantified using paired-end sequencing [165, 172]. The ChIA-PET method produces two types of ligation-products: self-ligations and inter-ligations. The self-ligations are caused by self-circularization ligation of the same DNA-fragment, resulting in the two sequencing pairs being placed very close together when they are mapped back to the reference genome. The self-ligations can therefore be used to determine the sites (or anchors) that are involved in chromatin interactions. The inter-ligations come from ligation between two different DNA-fragments, and are characterized by being placed further away from each other when mapped. The inter-ligations can then be subsequently used to determine the number of times each anchor was proximal by simply aggregating the contact frequencies between all anchors [173].

It is important to consider that ChIA-PET, unlike other 3C-based methods, will exclusively find interactions between regions that are bound on both sides by the same specific factor. Due to this requirement, it becomes difficult to determine if interactions are directly caused by the factor binding, since for example knock-down of the factor is impossible. Also, the data produced so far has been shown to have a low signal to noise ratio [165].

1.3 Computational techniques

Many of the newly developed molecular techniques described in the previous section have resulted in new types of genome-wide data sets, consisting of interaction frequencies between pairs of regions throughout the genome. Many of these new types of data have required new statistical and computational methods for filtering, quality checking, pre-processing and analysis. In this section, an in-depth discussion of some of these methods will be given. Since the focus of the thesis has been mainly on genome-wide methods such as Hi-C and ChIA-PET, methods relevant for these technologies will also be the focus of this section.

1.3.1 Hi-C data preprocessing

1.3.1.1 Mapping

The first step of Hi-C data analysis consists of mapping reads back to the reference genome. Since Hi-C interactions are determined using paired-end sequencing, mapping both ends of each paired sequence results in three different scenarios for each pair. Either none of the pairs are successfully mapped back, one of the two pairs are mapped, or both pairs are mapped back to the reference. However, an additional complication arises when considering that each of the ends can non-specifically map to several locations in the genome, due to repetitive and non-unique sequences in the genome. In the initial Hi-C paper [85], the authors chose to include such multi-mapped reads using the mapping software Maq, which has been speculated to cause some bias in repeat-associated regions in the genome [174]. In subsequent re-analysis of the original data, the mapping procedure was refined by the use of an iterative mapping scheme where reads were first truncated to a smaller length before genomic mapping (using Bowtie). Reads that were not aligned uniquely at both ends in the first round were then subsequently re-aligned by iteratively including larger portions of the reads. The end result was the set of read pairs with mapping status for both ends. The authors argued

that single-end mapped reads could be included in downstream analysis to avoid artefacts resulting from excluding reads that happen to interact frequently with repeat-regions for one of the ends. Reads that are not uniquely mapped at both sides should be discarded, according to the same authors [86].

1.3.1.2 Quality filtering

After mapping of reads, filtering out multiple technical artefacts is necessary. First, self-circularization or un-ligated (dangling-end) products will result in reads that map with both ends in the same restriction fragment, and should be removed. In addition, reads from neighboring fragments that face toward each other should be removed, as these can be the result of errors in the pull-down of the DNA fragments. Also, reads that map multiple times at the exact same location are often the result of biased PCR-amplification, and should also be removed [86, 175].

Due to the size-selection step in the Hi-C protocol, mapped reads are expected to be close to the restriction enzyme cut-site. However, occasionally, mapped reads will be found far away from these sites, due to for example physical breakage of chromatin (random breaks). The reads that are further away from the cut-site than what is expected from the size-selection, are therefore filtered out. Due to possible bias resulting from the size of the restriction fragments, very short and very long restriction fragments are filtered out. Imakaev et al. also recommended removing top 0.5% of fragments with the greatest number of reads, to avoid PCR artefacts. This cutoff may vary depending on the different data sets used, however [86, 175].

To account for differences in repeat-structure of the various restriction fragments, Yaffe & Tanay computed a mappability score for each fragment by artificially cutting the reference genome into partially overlapping 50 bp “reads”, and re-mapping them to the genome, thereby obtaining the mapping percentage for each fragment. The authors recommended removing the fragments with less than 50% mappability [87].

1.3.1.3 Binning and contact matrix generation

Even though the Hi-C data in theory provide contact frequencies between all-versus-all restriction fragments that remain after filtering, such a dataset would in practice be both very sparse (at current sequencing depths) and difficult to analyze in practice (with current computers). Therefore, the restriction fragments are grouped together into equally sized bins, and read-counts are aggregated across the bins, excluding reads within the same bin. The result will therefore be a symmetric matrix consisting of interaction frequencies between bins covering the entire genome. The bin-size chosen will naturally vary depending of the size of the genome under study, and the amounts of reads that remain after filtering. In the initial Hi-C paper, bin-sizes of 1Mb and 100kb were chosen. At subsequent studies, Dixon et al. have chosen bin-sizes as low as 40kb for the human and mouse genome, but these were only used for studying interaction frequencies at low genomic distance within chromosomes (along the diagonal of the matrix) where data are much more abundant [88]. See Figure 1.5 for examples of contact matrices at different resolutions.

1.3.1.4 Bias-correction and normalization

Several authors have pointed out that despite removal of technical artefacts at the level of individual reads or restriction fragments, other factors will cause additional biases at the level of bins. One of the most obvious biases is the differing frequency of restriction fragments within each bin, caused by the non-uniform distribution of restriction fragments throughout the genome. Such bias, in addition to biases caused by differences in GC-content, fragment length and mappability, was pointed out by Yaffe & Tanay [87]. They proposed a probabilistic procedure to model and thereby remove these specific biases from the initial contact maps. In their procedure, the authors model biases using a non-parametric step-function based on binning the GC-content and fragment length to model jointly the effect of the biases at the level of individual restriction fragments, giving the expected number of reads for a given pair of restriction fragments. The expected number of reads was then used to create an expectation matrix at the bin-level, which was subsequently used to cancel out the biases from the observed matrix [87, 175]. A similar method, called HiCNorm, has recently been proposed. In HiCNorm, a much simpler parametric model based on Poisson regression is applied at the level of bins, instead of restriction fragments. Due to the modeling of bins, HiCNorm is also much faster to compute [176].

Another correction method at the bin-level has been proposed by Imakaev et al. [86]. In that model, called iterative correction and eigenvector decomposition (ICE), the authors do not assume any specific biases. Instead, the authors argue that it is possible to factorize biases according to the difference in "visibility" between the bins, thereby incorporating all biases. The authors start by removing contacts of adjacent bins, and additionally, bins with the 2% lowest number of contacts, to avoid that these bins affect the correction procedure. The iterative correction procedure then consists of solving the following equation:

$$O_{ij} = b_i b_j T_{ij}, \tag{1.1}$$

where O_{ij} constitutes the observed matrix of contact frequencies, and the b_i and b_j factors are the biases for the bins involved in the contacts between bin i and bin j. T_{ij} is then the matrix of unbiased relative contact frequencies, which is defined such that each row and column of the upper triangular matrix sums to 1. The biases, and therefore T_{ij}, can be found by using an iterative procedure that converges to the solution of equation 1.1.

1.3.2 Domain identification

As discussed in section 1.1.3.4, mammalian genomes have been found to consist of domains at various resolution scales, such as A and B compartments using a bin-size of 1 megabase, and TADs using a bin-size of 40 kilobases. In this section, the methods for identification of such domains will briefly be discussed.

1.3.2.1 Principal component analysis for compartment analysis

To identify the A and B compartments, Lieberman-Aiden et al. used principal component analysis (PCA) by first dividing the observed Hi-C matrix by the expected matrix given by

the genome wide average interaction frequencies for the various genomic distances between bins. This matrix was then further processed by calculating a new matrix where each cell corresponded to the Pearson correlation between the row and column of that cell. The authors observed that the patterns in the resulting correlation matrix clearly consisted of two separate compartments. By performing PCA on this matrix, they found that the first principal component (in all but two chromosomes) gave positive values at what the authors called compartment A, and negative values for compartment B [85].

In Imakaev et al. [86], this procedure was further extended to incorporate the iterative correction procedure (ICE) explained in section 1.3.1.4. In that paper, the authors performed PCA on the unbiased relative contact matrix (T_{ij} in equation 1.1) instead of the correlation matrix, and by doing so they argued that relevant information can be found not only in the first principal component, but also in the second and third component.

1.3.2.2 Identification of TADs

One of the first genome-wide studies of topologically associating domains (TADs) was described in Dixon et al. [88] using a bin-size of 40kb (discussed in section 1.1.3.4). To identify TADs, the authors started by defining a directionality index (DI), defined for each 40kb-bin, using the following equation:

$$DI = \left(\frac{B-A}{|B-A|}\right)\left(\frac{(A-E)^2}{E} + \frac{(B-E)^2}{E}\right), \tag{1.2}$$

where A corresponds to the number of reads that map from the given bin to the upstream 2Mb region, and B is the corresponding number of reads that map downstream. E is the expected number of reads under the null hypothesis, and is given by $E = (A+B)/2$ (equally likely to go upstream and downstream). This index is based on the χ^2 test statistic, with the factor to the left giving the direction of the bias. The authors then trained a hidden markov model (HMM) based on a mixture of gaussian distributions, using three states corresponding to "upstream bias", "downstream bias" and "no bias" to determine the hidden sequence of states. The TADs were then found by scanning through the determined hidden series of states, such that the start of a TAD was defined by occurrence of a downstream bias state, and the end of the TAD was found at the last of a series of upstream bias states.

1.3.3 Building 3D models of chromosomes

With genome-wide chromosome conformation data such as Hi-C, the possibilities for building models of chromatin structure are emerging. Two broad strategies for building such models have been developed, both of which will be discussed here. The first strategy, referred to here as restraint-based structure determination, uses 3D interaction frequencies to determine an optimal structural model of the data. In the second strategy, the goal is to determine general organizational folding states of chromosomes, using statistical and physical principles. The goal of both these strategies is often to obtain an understanding of the physical properties of the chromatin structure and relate this structure to underlying functional properties to gain further insight into the biological processes that govern the regulatory programs of cells.

1.3.3.1 Restraint-based structure determination

In restraint-based modeling, the goal is to determine the structure that corresponds to a given set of restraints (or constraints), for example as given by a set of distances or contact points. In practice, these methods often assume that there is a consensus 3D structure across a population of cells, and that homologous chromosomes are structurally similar [175]. In these methods, the restraints are applied as forces between pairs of loci to enforce that the distances between the loci in the predicted structure resemble the original restraint-map. This is often done by minimizing a scoring function that incorporates several physical restraints, such as biophysical properties of the chromatin polymer chain, in addition to the distance-restraints. Several methods have been proposed for performing such modeling, both at local and global resolutions.

Optimization-based methods In the simplest models, only a set of interaction frequencies is used to determine the final structure. Fraser et al. [177] used 5C data from the HoxA cluster, and converted the interaction frequencies into spatial "distances" by simply taking the inverse of the interaction frequencies and optimizing a piecewise linear 3D curve by minimizing the difference between these "distances" and the Euclidean distances between points in the curve, using an iterative procedure moving points in small steps. This computational strategy was implemented into a program called 5C3D [177].

In a slightly more sophisticated approach, Bau et al. converted interaction frequencies from 4C in the yeast genome into distances based on principles in polymer packing. The authors then let the chromatin structure be represented as a series of beads for each restriction fragment. A bounding constraint corresponding to the nuclear space, and constraints to enforce chain-connectivity and to avoid bead-clashing were then used in combination with the distance constraints. The distances between the beads were then fitted to the observed (4C-based) distances while simultaneously enforcing all constraints. This method was used to propose a 3D structure for all chromosomes in yeast [98, 147, 178].

In another strategy, Di Stefano et al. used a steered molecular dynamics approach by using co-expression of genes on chromosome 19 as a proxy for 3D co-localization information. In that study, the authors used expression information to find pairs of genes that were significantly correlated across cell-types in terms of their expression. By assuming that significantly co-expressed genes were also co-localized in 3D, the authors were able to show that as many as 80% of the co-expressed genes could be co-localized simultaneously. In their setup, the chromatin fibre was modeled as a chain of beads with constrains corresponding to chain-connectivity, bending energy and repulsive forces, in addition to a co-localization constraint based on the significant gene pairs [71].

Bau et al. used the Integrative Modelling Platform (IMP) to infer structural ensembles based on 5C data on local regions. In this IMP-based method the authors converted the 5C data into a Z-score of log-values and applied distance-restraints proportional to the inverse of the Z-scores. In IMP, a particle is assigned to each restriction fragment with size proportional to the number of nucleotides in the fragment. A large number of restraints were added, such as clash-avoidance between particles, and a series of restraining oscillators (springs) between particle pairs. These springs were applied to ensure that distances between particles

were according to the (inverse) Z-scores and relationships between neighboring particles. The chain of particles were then optimized by searching for solutions that satisfied all of the restraints simultaneously in an optimal way. By application of the same procedure multiple times, and starting with a random structure, an ensemble of structures satisfying the restraints was obtained. This ensemble of structures could then be clustered to give a representation of some of the dynamics of the chromatin fibre, as given by the restraints. This method has for example been used to explore the 3D structure of the α-globin locus [178, 179]. Even though some variability of the underlying structure is allowed when using this approach, it is not necessarily the case that this variability reflects either the biological variability or the statistical uncertainty underlying the data [180].

Probabilistic modeling methods To allow for the introduction of statistical uncertainty, and a larger degree of dynamics in the structural models, methods aiming at describing a probability distribution of structures, instead of determining a single, optimal structure, have been proposed.

In one of the first such methods, called MCMC5C, [181] an initial structure consisting of a single chain of beads is iteratively changed using a random set of moves. By using a Monte Carlo Markov Chain (MCMC) procedure, the moves can be accepted or rejected depending on whether the new structure is more probable, given the data. Given a sufficient number of iterative steps, the resulting structures will be sampled from the posterior distribution. By running many such simulations in parallel, the resulting ensemble of structures is representative of the distribution of structures that fit the observed data. This model still relies on assuming a consensus structure underlying the observed data, however.

Another, similar, approach has recently been proposed, where an MCMC sampling procedure was also used [182]. In that study, the authors proposed two models, BACH and BACH-MIX. While BACH (Bayesian 3D constructor for Hi-C data) assumes a consensus structure, BACH-MIX assumes a mixture of structures. By applying both methods in parallel, the authors assessed which of the two models that described the structure best, thereby gaining insight into the degree of variability of the underlying structure [180].

Due to the high computational demands of such probability-based modelling, these methods have so far only been applied to relatively small systems, such as single chromosomes or smaller chromosomal regions.

1.3.3.2 Polymer models

Instead of sampling the 3D structure directly, other approaches to understanding the structure of chromatin based on biophysical principles has been used. Traditionally, the chromatin fibre has been understood in terms of the 30-nm fibre, as discussed in section 1.1.1. Recent evidence, however, does not fully support the view that chromatin above the 10-nm fiber is packaged in such a regular structure [183, 184]. The aim of understanding the overall folding principle of chromosomes has a long tradition within biophysics, and has typically been concerned with understanding different equilibrium states, and how these relate to actual observations of distances between loci in observed chromatin structure. The equilibrium globule state was long considered the most likely model of chromatin folding at larger scales [91,

185]. In this model, local parts of the chromatin fibre resemble a random walk, but the larger structure is affected by the confined space of the nucleus. The scaling of contact probability between two loci and the chain-distance (genomic distance) therefore has a characteristic scaling given by $s^{-3/2}$ (where s is the genomic distance), which reaches a plateau for larger genomic distances [186].

While these earlier models seemed consistent with the limited data from FISH experiments, recent analysis using Hi-C has shown that the scaling is closer to s^{-1}, with no plateau. Using simulations, this led Lieberman-Aiden et al. to conclude that chromatin was folded into a fractal globule, as opposed to the equilibrium globule. In the fractal globule model, chromatin consists of dense globules folded into further globular structures at higher levels, resembling a fractal curve. Importantly, fractal globule structures are free of knots, as opposed to equilibrium globules, which are highly knotted. The unknotted, local folding pattern of the fractal globule therefore represents an intriguing way for which the genome can arrange genomic elements at larger scales [186]. The current limitation of this model, however, is that it is formed during condensation, as opposed to decondensation, which is expected from mitotic chromosomes. However, recent studies have shown fractal globule like structures with scaling similar to s^{-1} also for simulations of decondensing chromosomes [71, 91, 187].

1.3.4 Hypothesis driven analysis of 3C-based data

With the recent introduction of high-throughput, genome-wide methods that couple next-generation sequencing with 3C-based methods, such as 5C, Hi-C and ChIA-PET, statistical methods and tools to find interactions that are biologically relevant among the many possible interactions between all considered regions have been developed. With conventional 3C, where 3D-interactions between two selected regions are assessed, analysis often simply consists of observing if interaction frequencies are higher than a control region. However, due to the increased number of random collisions between elements located close on the linear genome, it is important to control for such factors in the analysis. This could be achieved by comparing with control regions increasingly further apart, and to plot a curve of genomic distance versus interaction frequency. For significantly strong looping interactions, this curve is expected to show a bump of locally increased interaction frequencies around the specific loop [166].

With the introduction of the 4C method (discussed in section 1.2.4), by using microarrays, identification of significant interactions is performed by using a running mean over a set of probes, and for each set, comparing the running mean to the corresponding set for permuted data. The threshold level for which the difference between the permuted and the observed data is deemed significant is established by estimating the false discovery rate (FDR) for various thresholds, and set as the value where FDR < 0.05 [188].

One of the first statistical analyses of Hi-C data was presented by Botta et al. [108]. In that study, the authors investigated whether CTCF binding was associated with chromosomal interactions, as represented by the Hi-C data from Lieberman-Aiden et. al [85]. To do so, the investigators represented Hi-C interactions by using a multigraph, and started by asking if the number of contacts between pairs of nodes in the graph was higher than what would

be expected by a random graph. This was done by comparing the number of Hi-C interactions between all nodes in the observed graph with the average number of Hi-C interactions in graphs where edges were shuffled such that node-degrees were kept intact. By plotting the number of observed interactions in the non-randomized graph together with the number of interactions in the randomized graphs, the authors concluded that the interactions in the observed graph were non-randomly distributed. To investigate if such patterns could be related to the presence of CTCF, the authors then plotted the percentage of restriction fragments containing a CTCF versus increasing cutoffs of the number of interactions involved for each fragment. By observing that fragments with higher occurrence of interactions also had higher occurrence of CTCF, the authors concluded that CTCF-binding was related to increased occurrence of interactions, and speculated that CTCF could be a major organizer of chromatin structure [108].

With genome-wide 3D contact information both within and between chromosomes, as given for example by Hi-C, the possibilities for statistical analysis are manifold. Some of the types of statistical investigations that have been used are summarized in Figure 1.7, and will be discussed in detail in the subsequent sections.

1.3.4.1 Analysis of 3D co-localization of genomic elements

With Hi-C and similar genome-wide methods, it is possible to analyze the relative clustering of a set of genomic regions (see Figure 1.7). The first analysis of such 3D co-localization, was proposed by Duan et al. [98]. In that study, the authors proposed using the hypergeometric distribution to model the number of observed interactions k within the set of n selected genomic elements from the total of N genomic elements:

$$P(k|m, M, K) = \frac{\binom{m}{k}\binom{M-m}{K-k}}{\binom{M}{K}}, \tag{1.3}$$

where m is the number of possible interactions of the n selected genomic elements, and M is the number of possible interactions of all the N total genomic elements. K is the number of observed interactions between all the N elements. The underlying idea is to investigate if, from all possible realizable interactions, the observed k interactions between the selected genomic elements has a higher value than what would be expected from chance alone. The P-value in this case is then given by $1 - \sum_{x=0}^{k-1} P(x|m, M, K)$, which corresponds to Fisher's exact test.

To compute the number of observed and possible interactions, the authors started by considering a window of 5 kilobases around each position in the n genomic elements. To count the observed interactions, the authors did not consider the interaction frequencies themselves, but instead counted the number of times at least one interaction was found between any of the restriction fragments covered by the window and any of the other windows. The same strategy was used for counting the number of possible interactions, by considering all possible interactions between all restriction fragments inside the windows. Note that the authors only considered interchromosomal contacts.

A similar approach, also performed on interchromosomal contacts only, was used by Dai & Dai to investigate the 3D co-localization of transcription factor target genes [189].

In that study, the authors started by noting that the genomic elements in the selected set might have variable numbers of fragments, due to the non-uniform distribution of restriction enzyme fragments. The authors therefore chose to define an interaction between two genomic elements as present if at least one of the associated restriction fragments between the elements showed a contact. Additionally, the authors noted that if two elements are close to each other along the genome, an interaction between one of these elements and a third element might make it more likely for the other element to also contact this element. To control for such effects, the authors therefore chose to exclude all genomic elements closer than 60 kilobases. The significance of the 3D co-localization was then calculated in the same fashion as described previously, based on equation 1.3.

A different approach was taken by Ben-Elazar et al. [190], where the 3D co-localization of co-regulated genes was assessed. In that paper, the investigators used a ranking-based method to calculate the significance of intrachromosomal 3D co-localization taking into account genomic distance. To do so, they created two ranked lists of target genes for each transcription factor associated with the set of targets: One list was based on genomic distance and the other one was based on spatial (3D) distance. By using a method previously developed for ranked gene lists [191], the authors tested whether occurrence of co-regulated genes was more enriched at the top of the spatial list compared to the list based on genomic distance. This method was implemented and made available in a program called INSP3CT.

Yet another approach was used by Véron et al. [192], for analyzing whether evolutionary breakpoints where co-localized in 3D. To do so, they modeled (the logarithm of) the interaction frequency together with (the logarithm of) genomic distance in a generalized linear model, and introduced a boolean variable describing whether the pair of bins were part of a breakpoint pair. By using Analysis of Covariance (ANCOVA), the models were then compared to see if pairs with breakpoints were more co-localized than pairs without breakpoints. The authors then argued that, in addition to genomic distance, factors such as gene density and DNAse-I sensitivity (open/closed regions) could also affect the degree of 3D co-localization since breakpoints were more likely to be found in gene-rich and open chromatin regions. To take these factors into account, the authors included either gene-density or DNAse-I sensitivity as covariates in their model as well.

Note that the methods proposed in Ben-Elazar et al. [190] and Véron et al. [192] are slightly different from the other 3D co-localization methods, since in these two papers, 3D co-localization is characterized for a subset of selected interactions, instead of comparing all-versus-all interaction pairs for a given set of regions. Methods for this type of focused 3D co-localization analysis was also developed as part of **Paper II**.

Permutation tests Witten & Noble (senior author on the article described above [98]), noted that the assumption of the hypergeometric distribution as stated in equation 1.3 relied upon a faulty assumption of independency between the pairs of genomic elements [99]. They argued that the assumption of independency is faulty due to two reasons, as seen by considering that the tests are based on pairs of genomic elements: 1) given an interaction between the genomic elements i and j, and also an interaction between elements i and k, then the probability of observing an interaction also between j and k is higher. 2) If both the pairs i,j and i,k are part of the set, then by definition the pair j,k is also part of the set.

This observation led Witten & Noble to propose an alternative test, based on random permutation of the selected genomic elements, and comparing the interaction frequencies of the permuted set with the initial set. To do this, the authors selected randomly sampled sets of genomic elements of the same size as the original set, using uniform sampling. This randomization procedure was performed B times (with B set to 1000 in their studies). P-values were then computed as follows:

$$\frac{1}{B}\sum_{b=1}^{B} I(\frac{k^{*b}}{m^{*b}} \geq \frac{k}{m}), \tag{1.4}$$

where k^{*b} is the number of experimentally observed interchromosomal interactions among the genes in the random gene set, and m^{*b} is the number of possible interchromosomal interactions in the random set. $I(\cdot)$ is an indicator function returning 1 if the statement is true, and 0 otherwise, and k and m are the number of observed and expected interactions before randomization, as defined and described above.

A permutation strategy for assessment of 3D co-localization was also employed by Tanizawa et al., who combined 3C with next-generation sequencing to map genomic interactions in the fission yeast genome [193]. To calculate statistical significance of 3D co-localization of various sets of genomic elements, the authors randomly selected bins in the entire genome 1000 times, to build a null-model distribution, which was used to compare to the initial sum of interaction frequencies.

Engreitz et al. also performed permutation-based hypothesis testing in a study of 3D co-localization of chromosomal translocations [153]. To do so, the authors considered datasets of pairs of genomic positions involved in interchromosomal translocations. To investigate whether the pairs of translocations were significantly co-localized in 3D, as defined by Hi-C datasets, the authors started by computing the average, bias-corrected interaction frequencies between all pairs of interchromosomal translocation positions. Then, four different permutation strategies were performed: 1) randomly selecting regions of the same size on the same chromosome pairs, 2) fixing one of the regions, and randomly selecting a region of same size on the other chromosome, 3) fixing one of the regions, and randomly selecting a region of the same size on another chromosome, 4) fixing one region, and randomly selecting a different translocation region from the entire set of translocations. The four different permutation strategies were hypothesized to correct for different systematic differences, such as differences between chromosomes (1), predispositions of translocation sites to frequently interact on the same chromosome (2) or predispositions of translocation sites to frequently interact across the genome (3). Additionally, the authors controlled for the general propensity of regions of the same compartment to be involved in contacts (see section 1.1.3.4). This was done by permuting positions such that a randomly selected region was selected only within the same compartment (A or B).

In the paper by Véron et al. [192] discussed above, the authors argued that their proposed models made several assumptions about independence and variance, and therefore devised a permutation test to confirm their initial results. In their permutation test, the mean number of 3D interactions for the regions containing a breakpoint pair was compared to pairs not containing breakpoints, selected at random. To take into account factors such as genomic dis-

tance, gene-density and open/closed regions, the authors classified these factors into discrete classes, and maintained the discrete classes during the randomization.

Permutation tests have also been the basis for many of the developed methods in this thesis. Specifically, in **Paper I**, a permutation test for 3D co-localization analysis was developed, and in **Paper II**, this concept was expanded into a range of permutation tests for analyses of 3D co-localization in different settings.

Graph-based permutation tests Another permutation-based approach was proposed by Kruse et al. [194]. In that paper, the investigators represented Hi-C data as a network (or graph) with nodes corresponding to restriction fragments, and edges corresponding to observed 3D interactions between them. P-values were calculated in the same fashion as for equation 1.4, but instead of randomization of the positions of the genomic elements, they randomized the edges in the network. As noted by the authors, randomization (or rewiring) of edges must be done in such a way as to avoid introducing artificial biases into the network topology. The authors therefore proposed an edge-randomization strategy where node-degrees are kept intact, but where edges initially are randomly rewired. In the next step, the network is then corrected by adding triangle-edges (between node-triples) until the ratio of the number of observed triangles over possible triangles in the network matches the original network. Finally, edges are added to nodes that are neighbors on the chromosomes, again to match the original network. This procedure is repeated B times to obtain a P-value as given by 1.4. The authors also proposed to measure the extent of co-localization by calculating the deviation between the observed number of edges and the average number of edges, and dividing by the standard deviation (the Z-score).

Wang et al. also proposed a graph-based analysis approach of 3D co-localization, similar to the model in Kruse et al. [194]. However, instead of randomly rewiring edges in the graph, Wang et al. proposed to randomly select nodes instead, motivated by the fact that it is challenging to maintain topological structure of the original graph without introducing bias. The randomization of nodes is performed in a way dependent on what the nodes in the graph originally represent. For example, if nodes represent genes, a random set of nodes corresponding to genes is selected. However, if the input consists of genomic regions, then new random starting coordinates are selected, while maintaining the length of each region. Nodes are then selected by using the nodes that overlap the randomly positioned regions. P-values are estimated using a similar approach as in equation 1.4. All analyses are performed only on interchromosomal interactions [195].

1.3.4.2 Inference of significant interactions

Significant interactions in Hi-C Another type of analysis possible for Hi-C and similar data is the identification of significant interactions among all possible interactions. One of the first methods for doing so was presented by Duan et al. [98]. In that paper, the authors wanted to infer the interactions that were more frequent than the background signal, for both inter- and intrachromosomal interactions. To do so, they treated inter- and intrachromosomal interactions separately. To assign a P-value to each interaction, the authors started by assuming a uniform probability model for interactions between pairs of restriction fragments.

Specifically, for a total number of M possible inter-chromosomal pairs of restriction fragments, the probability (under this model) of observing any particular interaction is $p = 1/M$. The probability of observing k interactions is then given by the binomial distribution:

$$P(K = k) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{1.5}$$

where n is the total number of observed interchromosomal interactions. For intrachromosomal interactions, the authors argued that the genomic distance between restriction fragments needs to be taken into account, since the chromosomes act like a polymer where smaller genomic distances give higher probability of random contacts. To correct for this, the authors performed the test as given by equation 1.5 separately on interactions that were grouped according to genomic distances into bins of 5 kilobases each. Since P-values were calculated separately for each genomic distance bin, the authors noted that the P-value is conditioned on the genomic distance. Finally, since P-values were calculated for a large number of interactions, the authors performed multiple testing correction.

This method was refined in a follow up-study by Ay et al. [196]. In that study, the investigators sought to focus on the identification of significant interactions at the intermediate genomic distance scale (∼50kb-10Mb). To do so, the authors jointly modeled the genomic distance-dependent random looping effect and the various technical biases observed for Hi-C data (discussed in section 1.3.1.4). The authors started by noting that the binning of genomic distances as presented in [98] is problematic due to sharp transitions in the contact probability from one bin to the next. Therefore, the authors instead fitted the interaction frequencies using a monotonic spline fitting procedure to obtain a smooth estimate, $f(d)$, of the contact probability at given (exact) genomic distance d. The authors noted that outliers in the contact frequencies result in bias in the estimation of the spline, and therefore proposed estimating a refined spline by excluding outliers. To incorporate bias (see section 1.3.1.4) into the probability, the authors used the ICE procedure by Imakaev et al. [86] (discussed in section 1.3.1.4) based on the raw contact map in parallel with the spline fitting, to obtain estimates of the bias for each bin (b's in equation 1.1). The bins with very high or very low bias were removed, and the rest of the bins were used further. The corrected probability of a contact between bin i and j, was then defined as $p = f(d)b_i b_j$, which was used in equation 1.5 to obtain the probability.

A similar approach was used by Lin et al. in a study using Hi-C to study B cell development [197]. In that study, a background model of Hi-C interaction counts was generated by incorporating genomic distance and the number of contacts (sequencing depth) for each bin. To incorporate genomic distance, the average count for each genomic distance was calculated, representing the expected interaction frequency at various genomic distances. This expectation function was then combined with scaling factors that ensured that the interaction frequencies per bin were the same in the full expectation and in the observed data. To identify significant interactions, the authors then applied a binomial test similar to equation 1.5, but letting the number of possible interactions (M) be equal to the number of interactions involved for the bin with the fewest total interactions for each pair. This method was then implemented into a generic software suit for next-generation sequencing called HOMER [198].

In Jin et al., a slightly different approach was taken [199]. In that study, authors mapped chromatin interactions at extremely high depth using Hi-C in a human fibroblast cell-line (IMR90). By focusing on intrachromosomal interactions spanning genomic distances <2Mb, the authors were able to characterize interactions between individual restriction fragments rather than using the binning approach (see section 1.3.1.3). To identify significant interactions, the investigators started by fitting a model to estimate the expected signal at a given genomic distance, fragment length and GC content, in a model similar to the model by Yaffe & Tanay [87] described in section 1.3.1.4. This conditional expectancy was then used as a basis for fitting a negative binomial model to the interaction frequencies between the restriction fragments, giving a P-value for each pair of restriction fragments (within 2Mb). To finally determine the significance of interactions from a given restriction fragment to all neighboring fragments, the investigators used a peak-calling procedure, where peaks were defined as stretches of fragments with P-values < 0.1, in addition to requiring that at least one of the fragments had P-value <0.05 and that the total tag count > 10. The P-value was then recalculated for the whole peak, with the requirement that the P-value was < 0.005 and the tag count >15.

Significant interactions in 4C-seq A permutation approach has recently been proposed for significant interaction detection in 4C data also, in a method called fourSig [200]. In fourSig, initial data pre-processing is performed before significance of interactions from the viewpoint to each restriction fragment is assessed. To determine significance, fourSig randomly associates the reads with a specified window size of restriction fragments within the same chromosome. Based on a large number of repeated randomizations, the observed number of interactions between the restriction fragments is then compared to the distribution given by the randomizations, to give a P-value. For a given FDR-threshold specified by the user, the required number of interactions to reach significance is determined.

Another similar permutation strategy, where also the genomic distance between fragments is taken into account, starts by determining the relationship between genomic distance and interaction frequency, and calculating a Z-score based on the expectation and standard deviation conditional on the genomic distance. In a similar fashion as described above, the interactions are permuted to find the Z-score where the false discovery rate is below a given threshold [201]. A tool implementing this method in R/Bioconductor, called r3Cseq was recently published [202].

Significant interactions in 5C Detection of significant interactions is also a relevant problem for 5C data (discussed in section 1.2.5). In a study where interactions between promoters and distal sites were assessed for 1% of the genome (the ENCODE pilot regions), the authors estimated the background signal as given by the genomic distance and the bias for each restriction fragment, and found interactions significantly above the background signal. To correct for the bias per fragment, the authors computed the ratio of the average interchromosomal signal and the overall interchromosomal signal. Each interaction was then scaled by the product of the ratios of the two restriction fragments involved. After this correction, the signal from the genomic distance was estimated by fitting a smoothing spline to obtain the expected interaction frequency and the standard deviation for each genomic distance.

The corrected signal was then calculated as the Z-score, which was fitted to a Weibull distribution. Significant interactions were then identified by computing the P-value for each restriction fragment as compared to the background Weibull distribution [120].

Significant interactions in ChIA-PET Also for ChIA-PET data, the identification of significant interactions is a relevant problem. As mentioned in section 1.2.7, ChIA-PET typically involves two separate steps, one for identification of "anchors" where interactions occur, and a subsequent step where interactions between anchors are quantified. Like for Hi-C (and similar) data, the observed interaction counts need to be compared to a background model to infer the regions that are significantly interacting. To calculate the significance of such interactions, Li et al. proposed to use a hypergeometric distribution of the interaction frequency between anchors (n_{ij}) [173]:

$$P(n_{ij}|n_i, n_j, n) = \frac{\binom{n_i}{n_{ij}}\binom{2n-n_i}{n_j-n_{ij}}}{\binom{2n}{n_j}}, \tag{1.6}$$

where n_i and n_j are the number of end-points involved for anchor i and j, respectively. And n is the total number of interactions. The reason for conditioning on the marginal sums (n_i and n_j) is to correct for the varying propensity of the anchors to be involved in contacts. In other words, like for Hi-C data, the method needs to take into account the bias caused by some regions being more easily detected than other regions. Also, note that the model in equation 1.6 is a model over the end-points, and not the interactions themselves. The $2n$ factor is therefore used, since there will be a total of $2n$ end-points for n interactions. To calculate a P-value, the Fisher's exact test is used.

In **Paper III**, a hypothesis test of ChIA-PET data building on this model was developed. However, in **Paper III**, the method allows for taking into account the genomic distance in addition to the marginal sums (n_i, n_j) and the total number of interactions (n).

1.3.4.3 Differential interaction analysis

Another type of analysis possible for Hi-C and similar data is the comparison of two different treatments, to identify the significant differences between the Hi-C matrices. One of the first such analyses was performed by Rickman et al. [203]. In that study, the authors compared two prostate cancer cell-lines, one normal benign prostate epithelial cell line (RWPE1-GFP), and the same cells with induced over-expression of ERG (RWPE1-ERG). To identify significant differences in the Hi-C matrices between cell-lines, the investigators compared all pairs of (1 megabase) bins in RWPE1-GFP with the corresponding pairs of bins in RWPE1-ERG, by using the Fisher's exact test, taking into account the total number of interaction frequencies involving the bins for the given pairs. This was done for both intra- and interchromosomal interactions, and multiple test correction was performed to take into account the number of repeated tests.

Another strategy for identifying differential interactions between two Hi-C datasets was proposed by Dixon et al. [88], and applied to Hi-C data on mouse embryonic stem cells (mESC) and cortex. In that study, the authors applied the binomial distribution, and to test for significant differences, all possible pairs of bins (within 5 megabases) were considered.

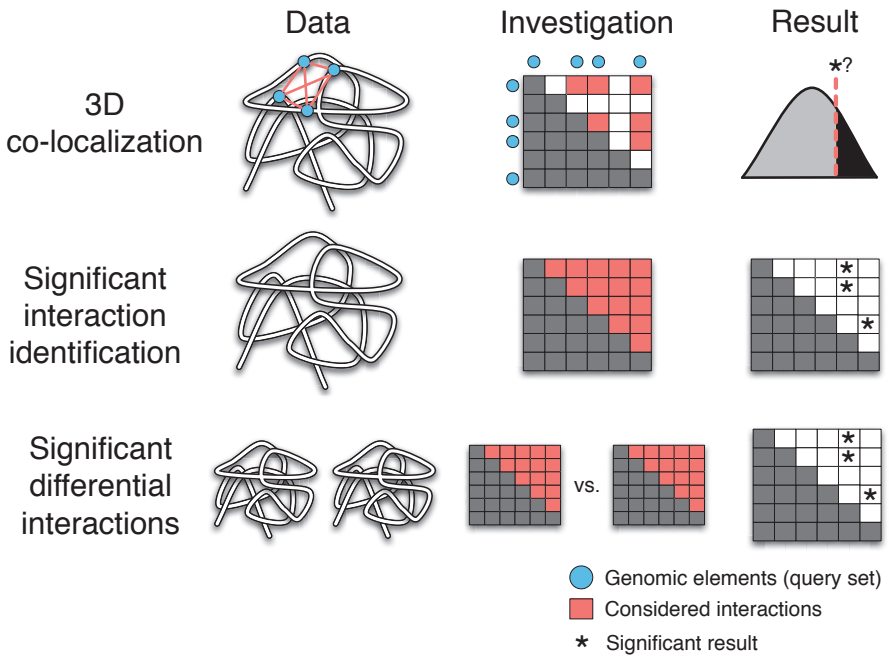

Figure 1.7: Three different classes of statistical investigations for 3C-based data. Top: 3D co-localization analysis, involving a pre-selected query set of genomic elements (blue circles). Middle: Identification of the pairs of genomic positions involved in statistically significant contacts. Bottom: Identification of significantly different interaction frequencies between pairs of genomic positions, by comparing between two different treatments.

Specifically, to test for a significant difference between a bin i and j comparing the two samples, the authors considered the total number of trials (n) to be equal to the sum of all reads between i and j in both treatments. To account for differences in coverage between the two samples, the authors considered the success-probability (p) to be the ratio of the sum of all interactions with the same genomic distance as between i and j in one of the samples to the total number of reads with the same genomic distance in both samples.

In the HOMER software suit (mentioned in section 1.3.4.2), identification of differences between treatments is made by first identifying significant interactions in one of the treatments, and then comparing the corresponding regions in the other treatments.

1.3.4.4 Correlation-based interactions

Another approach for identification of physical interactions between elements in the genome has arisen due to extensive mapping of epigenomic signals across tissues and cell lines in the ENCODE and Roadmap epigenomics projects [4, 5]. For example, by looking at DNAse I signal across 79 diverse cell types, Thurman et al. noticed that DNAse I hypersensitive sites (DHS) seemed to appear synchronously both at promoter and distal enhancer sites in a cell-type specific fashion [204]. Assuming that such correlations occurred due to physical interactions via transcription factors between promoter and distal sites, the authors then computed the Pearson correlation between each promoter site and all distal DHS within $\pm$ 500

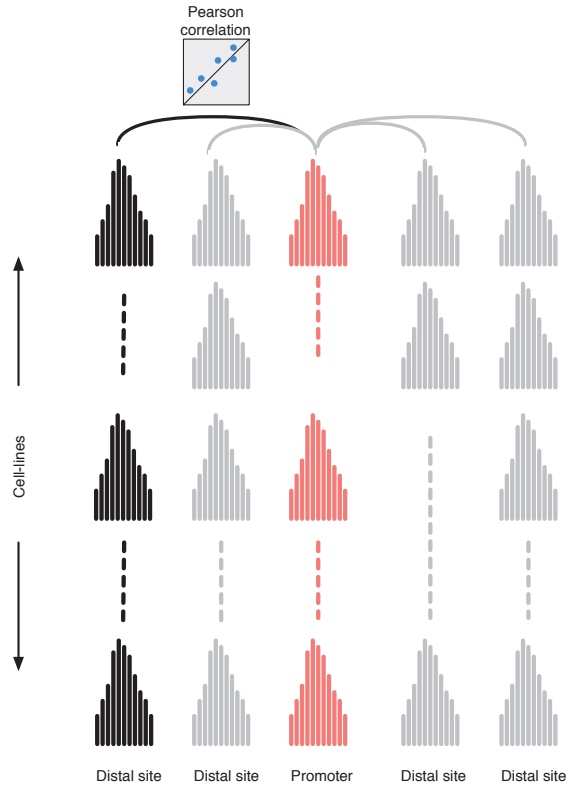

Figure 1.8: Using correlation to identify genomic interactions. By correlating signals such as DNAse I hypersensitivity between cell-lines from a promoter region (red) towards nearby peaks, significant correlations can be identified. Such correlations serve as putative interactions between distal regulatory sites and the corresponding promoter.

kb (see Figure 1.8). By selecting only highly significant correlations, and by requiring that the correlation > 0.7, the authors identified almost 600,000 distal DHS with significant correlation with at least one promoter. By comparing to 5C data, the putative correlation-based interactions were found to be enriched at sites with high number of 5C interactions.

A similar approach was used for detecting interactions in mouse, by calculating the correlation of polymerase II occupancy at promoters and H3K4me1 at enhancers across 19 different tissues and cell types. Using this approach, the authors found that correlations were enriched within compartments in the genome that overlapped with domains identified by Hi-C experiments. The authors were also able to validate a selection of the putative promoter-enhancer interactions by comparing to 3C data [205].

1.3.5 Descriptive and exploratory analysis

In addition to the more hypothesis driven methods discussed in section 1.3.4, most studies involving 5C, Hi-C, ChIA-PET and similar high-throughput chromosome conformation capture-data involves visualization and descriptive analyses, focusing on identifying the ef-

fect size of the interaction frequencies. One of the most basic approaches was presented by Lieberman-Aiden et al., where Hi-C data were divided into equally-sized bins and visualized using a heat map-representation of the underlying matrix (briefly discussed in section 1.2.6). In such heat maps, color intensity often represents the number of interactions observed for the given pair of bins (see Figure 1.5 for an example).

In this section, some of the common tools and methods used in the context of descriptive analysis of chromatin 3D structure will be discussed.

1.3.5.1 Contact enrichment analysis

Often, the raw visualization of contact maps will not reveal much of the underlying structure of the data, since the diagonal will dominate most of the displayed signal due the effect of local polymer looping (discussed extensively in section 1.3.4). Therefore, it is often necessary to display an enrichment of contacts relative to a background model. In a paper by Lieberman-Aiden et al., this approach was taken by estimating the expected number of interactions as given by the genomic distance, which was computed by calculating the average number of interactions for each genomic distance combination. The displayed heat map was then based on the observed matrix divided by this expected matrix. To further enhance the signal, the authors then created a new matrix where each cell in the matrix consisted of the Pearson correlation coefficient of the rows and column combinations of the observed/expected-matrix, which was subsequently displayed as a heat map [85].

A slightly more sophisticated model was devised in a study by van de Werken et al. [167], based on 4C-seq data. In that study, they performed two separate analyses for the regions around the viewpoint (which generally has much higher signals), and regions distal (or on different chromosomes) to the viewpoint. The background model was used to take into account factors such as fragment lengths and distance between restriction fragments. Results were then displayed as a log-ratio of the observed over the expected (background) signal, which was shown for different resolutions of bin-sizes.

In a paper by Sexton et al. [206], where Hi-C data in *Drosophila* was analyzed, a background model was calculated based on extending the model presented by Yaffe & Tanay [87] to also take into account genomic distance. This gave a model that estimated the probability of contact between two fragments, given fragment-lengths, GC-content and genomic distance. In a further extension, the authors also included domain information (hierarchical domain model). To visualize and quantify contact frequencies, the authors then computed the ratio between the observed and the expected (background) contact map.

Another, rather different, approach for estimating the frequency of interactions between features across domains in the Hi-C matrices is a method called Structure Interaction Matrix Analysis (SIMA) [197]. The goal of this method is to identify those genomic features that play a part in mediating interactions between domains. In this method, the total number of interactions between a given set of features found in different domains is computed. The total number of interactions observed between the features is then normalized to a background model taking for example genomic distance into account. To compute an enrichment score for each pair of features across domains, the normalized scores are compared to scores where the features are shuffled within the domains. The ratio of observed normalized interactions

divided by the average normalized interactions between the randomized set is then reported, for all combinations of features. With this approach, P-values can also be obtained by comparing the observed value to the distribution of randomized values.

In **Paper I**, an enrichment score for estimating the degree of over/under-representation of interaction frequencies, relative to an expected value, was developed in addition to a hypothesis test, in order to allow for quantification of 3D co-localization.

1.3.5.2 Visualization

Several tools for visualization of interactions and 3D information in the genome have been developed. The aim of most of these tools are to display interactions and allow for explorative analysis by comparing to other genomic datasets. In this section, some of the different methods for displaying and visualizing the outcome of 3C-based experiments will be discussed.

Circos One of the most widely used methods for visualizing significant contacts is Circos, a versatile comparative tool for displaying data in multiple ways [207]. In Circos, the entire genome (or individual regions) is displayed as a circle, together with positional information along the perimeter. Data are then displayed as tracks that are positioned relative to the circular reference coordinate system. Interactions can be displayed as arcs between the regions involved in interactions. Note that interactions when displayed like this are not restricted to 3D-interactions. Any relation between two or more positions in the genome can be drawn by the use of arcs (e.g. translocation events).

WashU Epigenome Browser A more specialized tool for visualization of epigenomic information, including three-dimensional interactions, is the WashU Epigenome Browser [208]. In this system, tracks such as histone modifications, transcriptional status, gene positions, compartments and interactions can be displayed simultaneously in an interactive browser environment. Hi-C tracks are displayed by using a tilted heat map representation, where the cells are colored according to the interaction frequencies. The tilted heat maps allow for easy visualization of local interaction hubs, such as TADs. ChIA-PET interactions are displayed as arcs connecting the various regions (see Figure 1.9).

HiTC HiTC is another visualization and data exploration tool, available as an R/Bioconductor package [209]. HiTC is capable of handling 5C and Hi-C data, and comes with standard processing tools, such as for quality control and normalization. For visualization, HiTC uses a similar heat map representation as the WashU Epigenome Browser described previously, and allows for re-binning and displaying at various resolutions. Genes and other one-dimensional tracks can also be displayed together with the 5C/Hi-C data.

3DGD 3DGD is a database and visualization tool for Hi-C data. This web-based tool allows for querying specific regions and displaying interaction frequencies from the query region towards a target region, using a histogram representation. Additionally, genes and protein binding information can be displayed together with the 3D data, to get an idea of spatial proximity between selected regions [210].

Genome3D Genome3D is a slightly different visualization tool compared to the other tools described here. Genome3D is a visualization tool and interactive viewing framework of chromatin 3D structure in three dimensions as xyz-coordinates. The engine allows for visualization of uploaded 3D structures, and allows interactive re-scaling of the same model, for viewing the whole structure in low resolution, or for zooming in at nucleosome scale. Positions in the structures can be annotated with additional data such as gene expression [211].

CytoHiC CytoHiC is a plugin for Cytoscape which allows for interactive viewing and comparing Hi-C datasets as networks, where the user specifies a set of genomic elements to be viewed. By specifying a set of bins, a contact matrix (Hi-C) and a set of genomic elements of interest, CytoHiC matches each of the elements to the corresponding bins and display the elements as nodes and the edges as the inverse contact frequency. A force directed layout algorithm is used to place nodes according to the inverse contact frequency, to give an idea of 3D positioning. Nodes can also be colored according to additional info added by the user, such as methylation or expression data [212].

ChIA-PET Tool ChIA-PET Tool is a software package for processing, filtering, mapping and analyzing ChIA-PET data [173]. A major feature in this software package is the ability to display ChIA-PET interactions in an interactive browser environment. This part of the pipeline is done in a separate browser system called G-Browse, which allows for viewing interactions together with annotated genes and other information, in an interactive fashion. Interactions are displayed as pairs of joined segments, and can be scaled out for a full-genome view visualizing all interactions within each chromosome at the same time.

1.3.6 Integrative chromatin analysis

Several computational tools have focused on analyzing 3D chromatin interactions in the context of other types of genomic and epigenomic datasets, with the goal of gaining insights into relationships between features, instead of analyzing each single feature separately. In this section, some of the tools used for integrative analysis involving spatial proximity between genomic elements will be described.

GWAS3D One of the first methods for coupling 3D chromatin data with DNA mutation data, was GWAS3D. In this web-based tool, information from a range of different epigenetic datasets have been collected, and merged into an integrated annotation and scoring pipeline. The goal of GWAS3D is to score a set of mutation variants from genome wide association studies (GWAS), by overlapping these sites with 3D information data (ChIA-PET, 5C and Hi-C) combined with epigenomic features, such as enhancer and insulator marks taken from various rich data sources. By combining these with information from cross-species conservation, sequence motifs and other features, a final prioritization score for each of the variants in the GWAS dataset is provided [213].

chroGPS An approach geared more towards visualization and association of different chromatin features was also recently proposed [214]. In chroGPS, epigenetic factors such as hi-

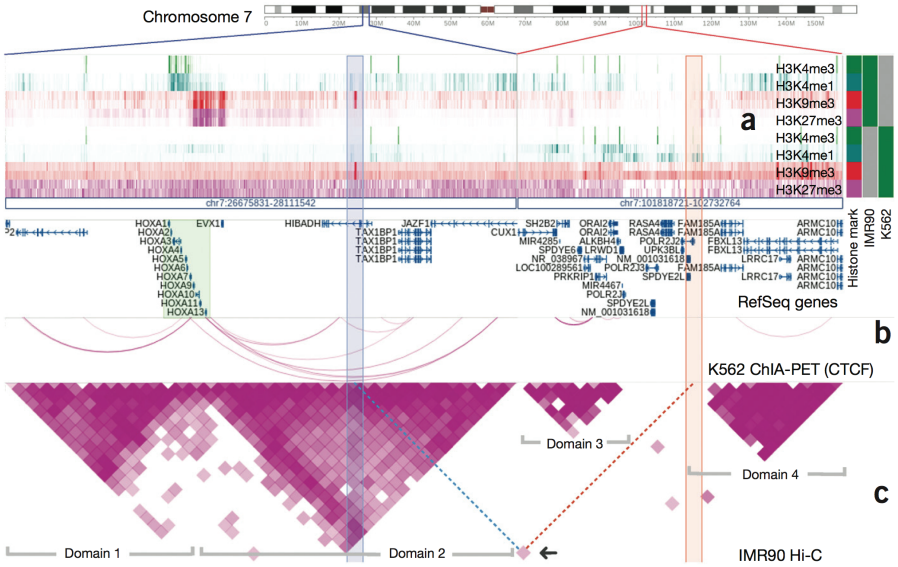

Figure 1.9: Illustration from the graphical user interface of the WashU Epigenome Browser. ChIA-PET interactions are illustrated using red arcs, while Hi-C domain architecture is illustrated using a tilted heat map representation along the diagonal. Reprinted by permission from Macmillan Publishers Ltd: Nature Methods [208], copyright (2014)

stone modifications, insulators and transcription factor binding are compared (all-versus-all) and clustered, using hierarchical clustering. The resulting similarities between the epigenomic features are then visualized using a 3D-map, where features are placed such that their proximity reflects the similarity between the features. Even though data such as Hi-C are not yet integrated into the system, the authors point to the possibility of including such information as well.

The Genomic HyperBrowser The Genomic HyperBrowser is a web server for integrative statistical analysis of relationships between diverse data types represented as tracks [215, 216]. The available functionality includes hypothesis testing, descriptive statistics, visualization and general processing of tracks, and allows for a range of genomic investigations. An illustrative example of an integrative analysis performed using this system, is analysis of relationships between chromatin states and disease [217, 218]. In **Paper I**, a basic 3D colocalization hypothesis test for Hi-C data was implemented into the Genomic HyperBrowser, and in **Paper II**, the functionality was expanded into a larger set of tools for general analysis of Hi-C data.

Chapter 2

Aims of the study

The overall aims of this thesis were to develop, implement and apply statistical and computational tools for analysis of genome-wide 3C-based datasets. The underlying goals can be summarized into four categories:

- Identify fundamental properties in 3C-based data, particularly Hi-C and ChIA-PET, to determine important factors to consider for valid statistical analysis.
- Develop rigorous statistical methodology for analyzing 3D co-localization of genomic elements in Hi-C data, taking such properties into account. A fundamental underpinning was the structured statistical concepts underlying the Genomic HyperBrowser, such as Monte Carlo permutation test functionality [215, 216].
- Implement a user-friendly tool based on the developed methods, building on the software components of the Genomic HyperBrowser [215, 216]. To this end, the structured representation of genomic tracks by means of the GTrack data representation format [219] should serve as a guiding principle underlying the developed methods. The implemented methods should be based on a flexible and expandable set of realistic null-models.
- Develop a statistical method for ChIA-PET data by taking into account fundamental underlying properties of such data.

Chapter 3

Summary of the papers

3.1 Paper I

With the recent coupling of next-generation sequencing to chromosome conformation capture-techniques, methods such as Hi-C have been developed [85], giving genome-wide interaction frequencies both within and between chromosomes. With such novel data, new statistical and computational techniques for data analysis are needed. One possible investigation with these data is the analysis of 3D co-localization, where the spatial proximity of a set of genomic elements is compared to an expected degree of co-localization under various assumptions.

In **Paper I**, our overall goal was to develop and apply a method for analysis of 3D co-localization for both intra- and interchromosomal interactions in mammalian Hi-C datasets. We started by considering the set of properties underlying such data that could possibly influence the statistical analyses. By considering these properties, we quickly came to the same conclusion as Witten & Noble, namely that obtaining a complete description of the null model in the form of a distribution function is difficult, and therefore a permutation test is probably needed. We also realized that the permutations had to be performed on the genomic positions, and not on the data itself, due to the high degree of transitive properties in such data [99].

As has been noted for 3C-based data previously, a fundamental property of such data is the increased chance of contacts for short genomic (sequence-based) distances. Since few prior methods had been designed for analysis of 3D co-localization within and between chromosomes for Hi-C data, we were motivated by taking this property into account. In **Paper I**, we proposed to use an approach where the expected interaction frequency and standard deviation, as given by the genomic distance, was corrected for as part of the test-statistic.

We also noticed that even though permutations were performed on the genomic positions, the measure of 3D co-localization would be affected by correlations between interaction frequencies, caused by the genomic distances between the elements of the query set. One of our working hypotheses was that interactions that were close to each other along the genomic sequence would have more similar interaction frequencies, even after correcting for genomic distance, than interactions far away from each other. We noticed that this property would result in correlation structure that could possibly influence the distribution of our test statistic. Therefore, to investigate the occurrence of correlations between interactions based on genomic distance, we computed the similarity between interaction frequencies for various

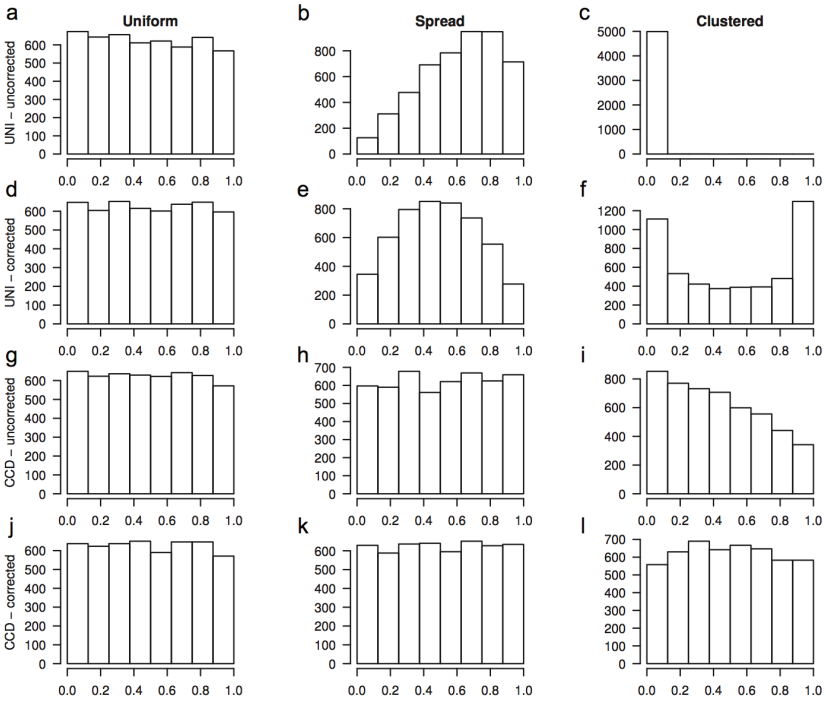

Figure 3.1: Histograms showing the distribution of P-values where elements of the query set are sampled uniformly along the chromosome (left column), spread out (middle column) or clustered (right column). Each row corresponds to different methods used for calculating P-values, with increasing degree of sophistication. The method proposed in **Paper I** is shown in the bottom row. This figure is based on intrachromosomal contacts, and the figure is reprinted from the Supplementary Material of **Paper I**.

pairs of bins in two recently published Hi-C datasets at two different resolutions. The results showed that, indeed as we expected, interaction frequencies between closely spaced pairs of bins were found to be highly similar, even at genomic distances of several megabases. These results motivated us to design a permutation strategy that would take such properties into account. We ended up with a permutation strategy we called Conserved Consecutive Distances (CCD), where the genomic distances between all consecutive pairs of genomic elements are conserved while being permuted, but higher-order distances are not conserved. To validate this approach, we needed to simulate 3D structures from a null-distribution where distances between randomly selected regions were not closer or further away from each other than expected by chance. We therefore simulated "pseudo-chromosomes" based on random walks inside a bounding sphere. Using our proposed permutation method, we evaluated 3D co-localization of query sets with elements being either clustered, spread or uniformly selected along the chromosomes. By comparing the CCD-strategy with a simpler strategy, where elements were randomly (uniformly) permuted, and where we either included or excluded the correction for genomic distance, we were able to show that the CCD-strategy were the only method that gave valid results even when the genomic elements of the query set were clustered together or spread out (see Figure 3.1).

Since recent papers had pointed out that the genome is partitioned into domains with differing degree of 3D interaction frequencies, and that regional preferences dependent on the relative positioning along chromosome arms influenced the data [85, 87, 88], we also explored such properties in our method as well. To take such properties into account, we devised the strategy of simply permuting (using CCD) within domains, and showed that this strategy seemed to allow for more specific hypothesis tests, where various properties can be included into the model depending on the particular research question.

In addition to investigating the significance of 3D co-localization by using P-values, we also implemented a quantitative measure giving the degree of enrichment of 3D co-localization, realizing that hypothesis testing on its own is often not enough to truly obtain biologically meaningful results. Interestingly, since we applied a correction for genomic distance to our test-statistic in order to analyze intrachromosomal as well as interchromosomal interaction frequencies, we could not simply use the test-statistic as a measure of enrichment. Instead, we devised a measure based on calculating the observed over the expected contact frequency, as given by the genomic distance, in addition to other properties such as compartments or relative positioning along chromosome arms. We did this by using permutations, similar to when calculating the P-value, and by estimating the expected contact frequency coming from genomic distance and all other factors, separately.

Finally, to also investigate the power of the method, we applied our method to several Hi-C datasets. We started by selecting data where we expected a high degree of 3D co-localization, such as promoter and enhancer regions, and confirmed that both the enrichment score and hypothesis test gave results as expected.

We then sought to investigate the 3D co-localization of regions involved in somatic mutations in leukemia cells, motivated by recent findings that had shown a link between mutational events and genomic 3D architecture [156]. By applying both the enrichment score and the hypothesis test on a dataset of somatic mutations in leukemia cells using three different permutation strategies, four different bin resolution sizes, and for intra- and interchromosomal data both separately and jointly, we found a statistically significant 3D co-localization of somatic mutations for intrachromosomal interactions. However, by inspecting the enrichment score, we found that the degree of 3D co-localization was marginal, and we found it difficult to conclude on the biological relevance of this result.

3.2 Paper II

In **Paper II**, the motivation was to develop a versatile and expandable web-based tool for analysis of 3D co-localization and other types of 3D analyses. We did this by expanding on the architecture of the Genomic HyperBrowser [215, 216], and by adding functionality in both data representation and analysis framework, into a tool we called HiBrowse. Since one of the main results from **Paper I** was that a permutation test was needed in order to get valid hypothesis testing of 3D co-localization, we re-used parts of the functionality for Monte Carlo permutation testing already implemented in the Genomic HyperBrowser.

Additionally, we expanded the method developed in **Paper I**, by considering that other types of 3D co-localization analyses were possible besides the all-versus-all nature of 3D co-

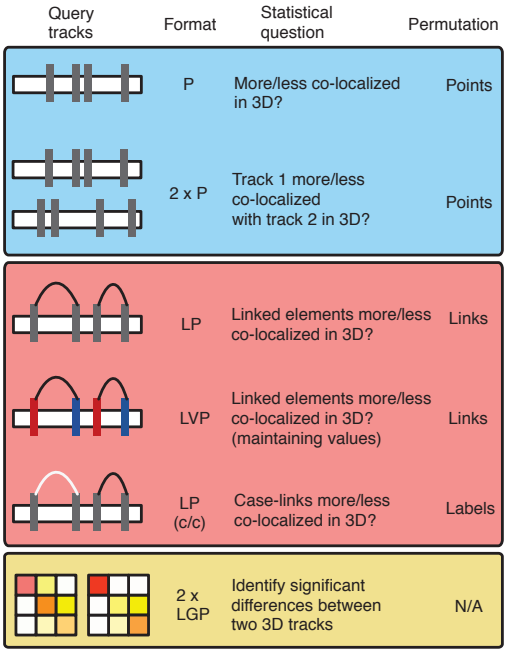

Figure 3.2: Three classes of analyses implemented in HiBrowse. The classes are defined according to the nature of the query track. With point tracks (P), all-versus-all 3D co-localization is used, and permutation is therefore performed on the points themselves (blue). When linked query track types such as linked points (LP) or linked valued points (LVP) are used, only the interactions defined by the links are considered, and permutations are performed while maintaining positions and shuffling of links. With 'case' and 'control' labels on the links, only links marked by 'case' are considered, and shuffling of labels is performed instead (red). Finally, identification of significant differences between two Hi-C tracks is possible, based on statistics developed in Robinson et al. [220] (yellow). The figure is reprinted and adapted from **Paper II**.

localization as presented in that paper. For example, we wanted a simple method to analyze the 3D co-localization of a selected set of genomic interactions. Motivated by the structured way in which statistical investigations are handled in the Genomic HyperBrowser system, we developed a set of statistical tools where the structure of the query dataset defines the types of statistical questions possible (see Figure 3.2).

To represent the possible query set types, we needed a versatile and generic representation of the underlying data types. To this end, we based the statistical method setup on the GTrack format representation [219], developed for the Genomic HyperBrowser, and capable of handling interconnections between elements. In GTrack, genomic data types are defined according to 15 different track categories dependent on four core informational properties of such data: gaps, lengths, values and interconnections. For example, a regular BED file of genomic elements, belong to the segment (S) category, and a single position (no length is defined) belongs to the Point (P) category. Each genomic element could also be accompanied by a value of some sort, and if so, the data belongs to the Valued Point (VP) or Valued Segments (VS) category. By allowing for interconnections between points, the tracks belong to the Linked Points (LP) category, or Linked Valued Points (LVP) if a value for each element is also considered.

In HiBrowse, when a 3D track such as a Hi-C dataset is selected, the second (query) track selected defines the possible types of analyses. We defined hypothesis tests according to three broad categories dependent on the type of the query track. The first category of statistical analyses was defined by considering query tracks of type P. When one track of type P is used, we have the same statistical analysis as presented in **Paper I**. When two tracks of type P are selected, however, we allow for considering the 3D co-localization between all possible pairs of interactions between the two tracks, but not within. In other words, it is possible to analyze whether genomic elements in one of the tracks are more (or less) co-localized with the elements of the other track (in 3D), than expected by chance.

The second type of statistical investigation arises when selecting a linked track, such as LP, LVP or LP with case/control categories associated with each link. In such analyses, only the interaction frequencies between the regions defined according to the links are considered. This allows for analysis of whether the linked elements are more (or less) co-localized in 3D, than expected by chance. In these types of analyses, permutations are performed on the links (or case/control labels), preserving the positions of all genomic elements in the query set, in contrast to the first type of statistical question.

The third type of statistical investigation is fundamentally different from the two first. Instead of analyzing 3D co-localization between a set of selected elements, this type of analysis identifies significant differences between two Hi-C (or similar) tracks. We realized that this type of analysis resembles the type of analysis used for detection of significant differences in digital gene expression data. Since the development of statistical tools for gene expression analysis has been rather extensive, we applied and implemented the same type of test as developed in Robinson et al. [220], designed for differential signal analysis with genome-scale count data.

The graphical user interface (GUI) of HiBrowse is based on Galaxy [221] and is implemented in integration with the Genomic HyperBrowser system [215, 216] (see Figure 3.3). The overall concept is to allow for easy access and user-friendly functionality for analysis of Hi-C and related data types in a range of settings, focusing on 3D co-localization analysis. In addition, some visualization and explorative tools are provided.

3.3 Paper III

In **Paper III**, we applied the acquired knowledge of topological properties of chromatin data from **Paper I** and the statistical analysis types of **Paper II**, to develop a statistical test of ChIA-PET interaction data. This type of statistical analysis is fundamentally different from the types of analyses considered in **Paper I** and **Paper II**, since the goal of this paper was to develop a model for identification of the individual interactions that are significantly higher than some background model (see Figure 1.7). For this type of question, correlation between elements caused by for example linear genomic proximity is not as important as for analyses of 3D co-localization, since the sum of interaction frequencies is not considered. However, properties related to the propensity of certain regions to be more interactive than others (such as compartments or domains) will still be important.

In essence, only one type of statistical test, based on the hypergeometric distribution, had

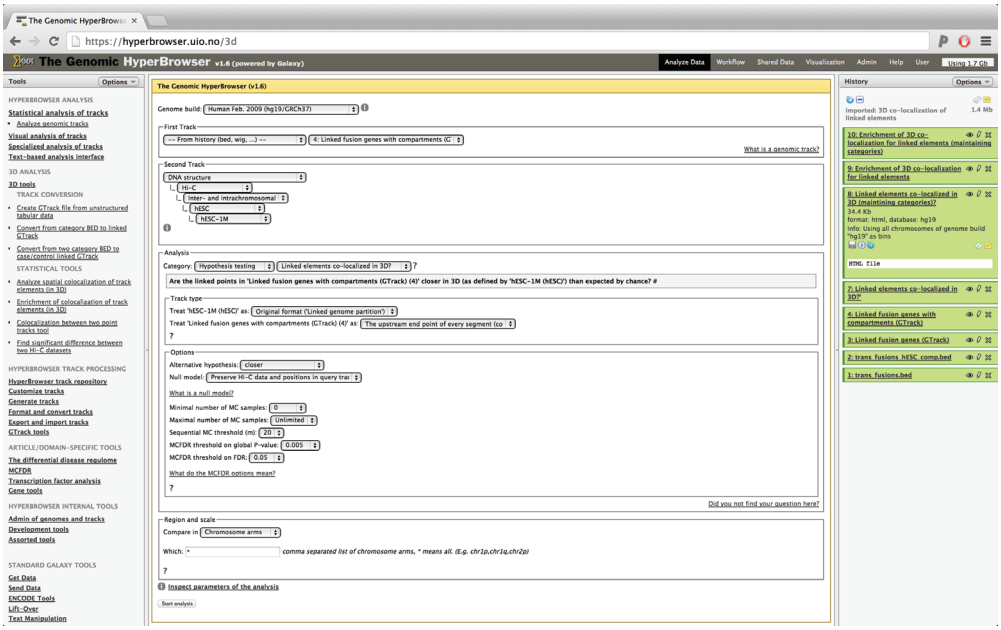

Figure 3.3: A screenshot showing the graphical user interface (GUI) of HiBrowse, based on the Galaxy and Genomic HyperBrowser system. The left panel displays the different types of analyses available in the system, the middle panel shows the work area, where options relating to the selected tool are displayed. By clicking the "start analysis" button, the selected analysis will be submitted to the server and will immediately appear in the right panel, where the final results together with all uploaded datasets are displayed. This particular example shows options related to an analysis of fusion transcripts defined according to linked elements.

been previously applied to such type of analyses [173]. While this model takes into account the degree to which pairs of regions are involved in contacts, the genomic distance between elements is not taken into account. We therefore decided to use the non-central hypergeometric distribution, (NCHG) which is a generalization of the hypergeometric distribution, allowing for taking the variable probability of interactions at different genomic distances into account.

Since no studies had previously been investigating the effect of genomic distance between anchor regions in ChIA-PET data, we started by comparing this effect on two different model assumptions. By simulating data that were both dependent and not dependent on genomic distance, we evaluated the effect of not taking genomic distance into account, in a similar fashion as was done in **Paper I**, by inspecting the distribution of P-values under the null-model. We found that not taking genomic distance into account (using Fisher's exact test) caused skewed distribution of P-values under the null-model. Using the non-central hypergeometric test, however, P-values were found to be uniform, as expected.

We then applied our method, and Fisher's exact test, to two publicly available ChIA-PET datasets, from Mcf7 and K562 cells. By comparing the number of significant interactions after multiple testing correction, we found that Fisher's exact test reported a very high number of significant interactions compared to the NCHG test. We showed that this was due to the fact that Fisher's exact test reported many significant interactions with short genomic distances,

caused by the higher number of contacts not taken into account in the model.

Finally, by showing several examples of regions with previously identified interactions, such as for the α-globin locus, we showed that the NCHG test allows for accurate detection of significant interactions without over-estimating the significance of short, nearby interactions.

Chapter 4

Discussion

In this PhD project, the main focus has been on inferential analysis of genomic 3D organization. The term "inferential" is here meant in a broad sense, as a way of analyzing data with the goal of drawing conclusions of some sort. Determining causal relationships (causal inference) is not considered here, as this requires specific study designs prior to analysis, for example involving temporal data collection. Some of the tools developed, for example as part of **Paper II**, are nevertheless more descriptive than inferential. Together, the developed tools should be considered broadly as tools for "data analysis" [222].

Specifically, the major goals of this project have been to develop, implement and make available relevant tools for analysis of 3C-based data, particularly genome-wide methods such as Hi-C and ChIA-PET. In the developed methods, much effort has been put into taking into account factors that are inherent in such data, and that distinguish them from other genome-wide datasets. To do so has required utilization of expertise from rather diverse disciplines such as statistics, informatics and biology, and coordinating these efforts into a usable result. Therefore, a high-degree of multidisciplinarity has been necessary to make this project work. Such multidisciplinarity, while having been rewarding for those involved, inevitably has come with a cost of being unable to delve deeply into a particular problem and learning all there is to it. Some of the major challenges to this project have therefore been to find the balance between usable solutions that still are theoretically sound. Consequently, the major part of the work has been a theoretical and methodological endeavor, with few direct biological results. The aim has been to make the methods usable through implemented software, with the hope that the methods can be extended and built upon in the future. Since results from Hi-C, and similar genome-wide 3C-based techniques, constitute a completely new type of genomic data, the challenge has also been to truly get a grip of the fundamental properties, biases and noise profiles of such datasets.

In this section, some of the challenges and general considerations that were encountered during the project will be discussed and put into context. At the end, some speculations as to what may be expected in the future, will be given.

4.1 Data quality and availability

The methods and tools developed during this PhD project have all been entirely dependent on the availability of public data resources, as no data production has been conducted in-house. Therefore, a large part of the project has been to scrutinize and thoroughly examine available data. During the development of the methods presented in **Paper I**, it became clear that Hi-C data in the form of raw contact frequencies were highly affected by technical biases that needed to be taken into account for proper analysis of such data [86, 87]. This is illustrative of the extremely unfinished state that these types of technologies were, and probably still are, in. Luckily, we were able to adapt and adjust our methods accordingly, with little extra effort necessary. This was probably possible due to the decision on employing a Monte Carlo based method, where no assumption on an explicit distribution for the data was needed. If similar adjustments to data happen in the future, Monte Carlo methods are likely to be a good choice, due to this very reason. Currently, there seems to be two ways of handling biases: 1) Biases are removed prior to analysis, and bias-removal is seen as a pre-processing step, 2) biases are incorporated into the model itself and treated as parameters that are adjusted for. In our methods, we have chosen strategy (1), which also seems to be the most common choice. In practice, we have done this by utilizing the method of Imakaev et al. [86] on all the data as an initial step, before even loading them into the data repository of our system. It is important to note that several other bias removal methods have been proposed (see section 1.3.1.4). We chose the method of Imakaev et al. [86] because it is implemented in a fast and readily available tool for easy processing of large amounts of data. The choice of bias-removal method is probably not crucial, as it has been shown that these methods in fact produce very similar results [86]. It is, however, uncertain at this point whether the bias removal procedures only remove bias, or if some biological signal could be affected as well. Since the performance of these methods are usually evaluated based on their ability to enhance reproducibility between biological replicates, it can be difficult to assess the exact effect of bias-removal in all circumstances. The use of e.g. FISH data for evaluation of these methods have been proposed as a viable alternative to shed light on these matters [175].

At the moment, we have chosen to make available in our system (as described in **Paper II**) a total of 14 pre-processed datasets at various resolutions in four different species (see Table 4.1). Inevitably, some datasets have not yet been added to the system, such as data for fission yeast [193], budding yeast [98], or the recently mapped 3D genome of the protozoan parasite *Plasmodium falciparum* [223]. We also have not included a newly released data set of single-cell Hi-C [149], or some newly mapped human cell lines [150].

In **Paper I**, the data used came from two separate sources. The data for investigating correlations between linearly proximal bins, and for assessing the power of the method, came from Dixon et al. [88]. We chose this dataset because it had the highest resolution (deepest sequencing) of the publicly available data sets. To study the 3D co-localization of mutational events, however, we used a cell-line more similar to the source cells of the mutation data, namely K562 from [85], even though this dataset had lower resolution.

A similar choice was made when choosing appropriate datasets for testing the methods developed as part of **Paper III**. We tested our method on a dataset from Li et al. [130], because it was of very high resolution, and since it had been studied extensively in the same

Table 4.1: Overview of installed and pre-processed Hi-C data available in HiBrowse (reprinted from Supplementary Material in **Paper II**).

Species	Cell-line/tissue	Treatment	Bin-size(s)	Ref
H. sapiens	GM06990		100k, 200k, 500k, 1M	[85]
H. sapiens	K562		100k, 200k, 500k, 1M	[85]
H. sapiens	GM12878		100k, 200k, 500k, 1M	[148]
H. sapiens	hESC		100k, 200k, 500k, 1M	[88]
H. sapiens	IMR90		100k, 200k, 500k, 1M	[88]
H. sapiens	RWPE1	ERG	200k, 500k, 1M	[203]
H. sapiens	RWPE1	GFP	200k, 500k, 1M	[203]
M. musculus	mESC		100k, 200k, 500k, 1M	[88]
M. musculus	cortex		100k, 200k, 500k, 1M	[88]
M. musculus	pre-pro-B		100k, 200k, 500k, 1M	[197]
M. musculus	pro-B		100k, 200k, 500k, 1M	[197]
D. melanogaster	embryo		10k,20k,40k,80k,160k	[206]
A. thaliana	Col	WT	200k,500k,1M	[224]
A. thaliana	Col	atmorc6-1	200k,500k,1M	[224]

paper. Importantly, this dataset was based on interactions involving RNA Polymerase II (RNAP II), allowing for interrogating 3D interactions involving promoters and enhancers, and therefore making it more easy to compare with previously mapped chromatin interactions. Other choices of datasets could have been possible, including data based on CTCF [110], H3K4me2 [70] or estrogen-receptor-α [171]. It is important to note that there is a potentially large difference between ChIA-PET datasets based on factors such as RNAP II, compared to factors that bind more specifically, such as CTCF or transcription factors. The broader binding patterns of RNAP II results in larger anchor regions for which interactions are aggregated. However, the exact effect of the size of the anchor regions has not been studied in detail. Further comparative analyses of such effects may therefore be needed to shed light on possible differences between broad and narrow anchor regions in ChIA-PET data.

Another important issue to consider, particularly for Hi-C data, is the fact that translocations and rearrangements in the samples analyzed, could cause bias in the resulting contact matrices, since regions linearly proximal in the genome will necessarily have higher number of contacts (mentioned in e.g. [86, 153]). Translocations, for example, may result in what appears as a high number of interactions between regions on separate chromosomes when reads are mapped to a reference genome without such rearrangements. With 3D co-localization analysis, where the contact matrix itself will be used as part of the null model, this is likely not to be so severe. However, for analyses where the goal is to identify significant interactions, or for differential analysis between two samples, this issue may be more important. A combined approach where Hi-C data is first used to detect translocations and re-arrangements, to either mask out or take such events into account in the model, and then used for statistical analysis, may be possible in the future. A similar approach could perhaps be used for ChIA-PET data as well, prior to methods such as the one presented in **Paper III**, but it is not clear how the uneven distribution of anchor regions in such data would be handled.

Copy-number variation specific to the interrogated cell-material could similarly lead to over- or underestimation of the number of interactions involved for such sites. Bias-correction methods could partially alleviate such effects, but this has not been studied extensively.

4.2 Implementational issues (Paper II)

Currently, few generic software suits exist to analyze Hi-C and similar type of data, and therefore, most of the functionality had to be developed from scratch. However, by building on the Genomic HyperBrowser, we were able to utilize the general statistical framework and the theoretical mindset already laid out through several years of research in statistical genomics. We note that the HiBrowse tool presented in **Paper II** is meant as a tool to be expanded as further methods are developed. In other words, the currently implemented tools are not meant as a final set of possible tools for Hi-C analysis. For this reason, the main scientific contribution in the HiBrowse tool is the flexible framework and the possibility for doing analyses easily on large datasets in a user-friendly environment.

One of the major challenges in making versatile and open-ended methods for Hi-C data analysis, such as the one presented in **Paper II**, is the size and complexity of genome-wide interaction frequency data. One of our goals with this project was to create a system where new statistics could easily be added, without having to do too many additional adjustments to the code. We solved this by applying a modular approach, where many of the common steps in the data analysis could be re-used for the different types of statistics. One example is the computation of the standardized interaction frequency for each genomic distance, which is a time-consuming computation resulting in a step function giving the expectation and the standard deviation for each genomic distance in a given Hi-C dataset. We noticed that these computations were only needed one time for each dataset at each possible resolution. Therefore, we implemented a caching strategy that saves the resulting step functions on disk after it is calculated for the first time, and then reuses these in all the steps where they are needed.

However, the computation time required for a large analysis in HiBrowse is, at the moment, quite high. Depending on the type of analysis, a regular 3D co-localization analysis of thousands of genomic positions can take several hours. The main reason for this is that the permutations of the positions during the Monte Carlo simulations require intersection of the permuted positions and the Hi-C bins at each step. In principle, it is easy to improve upon the computational time needed for these analyses, since for example results from individual chromosomes can be combined in the global result utilizing MapReduce functionality already implemented in the Genomic HyperBrowser system. However, it has proven difficult to implement such functionality for Hi-C data, since results must be combined from pairs of chromosomes. In addition to computational costs, 3D co-localization (and similar) analyses can be highly dependent on memory usage. To allow for fast computations, HiBrowse currently uses vector operations on objects such as large matrices in memory. The memory usage is again dependent on the size of the analysis (size of the query set), and using MapReduce functionality would improve on these matters as well. Problems relating to computation time and memory-requirements are likely to escalate drastically in the future, as the throughput from next-generation sequencing technologies is constantly increasing. For Hi-C

and similar data, this will result in better resolution in the form of smaller bin-sizes, resulting in drastically increased size of the arrays or matrices holding the data. It is important to note, however, that it may not be necessary to use a similar bin-size for all possible interactions. For example, since interchromosomal interactions are much more sparse, it could be possible to use a much lower resolution (larger bin-size) for interactions between chromosomes. Similarly, interactions in close proximity along the genomic sequence are known to be of much higher resolution than long-distance interactions. To take such differences into account, a more sophisticated data structure than a simple matrix is probably required, however.

Since the field of Hi-C analysis is so new, currently, no standard data format exists for representing Hi-C or other 3D interaction data sets. We therefore had to devise a unified way to represent such data, resulting in the GTrack format, recently published by several people in our group [219]. In GTrack, the genomic data are represented in a uniform way, such that diverse types of genomic datasets can be represented similarly. In **Paper II**, one of the main features needed for representation and analysis of data was a way of allowing for linked elements. This was needed in order to specify particular interactions that were to be considered for 3D co-localization analysis, similar to the types of analyses presented by Engreitz et al. [153].

We have chosen to build the HiBrowse system on the general framework of the Genomic HyperBrowser, which again is based on the Galaxy framework [221]. There are several advantages to using such a framework, instead of building a new framework from scratch. Since Galaxy is currently used by thousands of scientists world-wide, the graphical user interface (GUI) is well-familiar to many researchers. The use of methods in the Genomic HyperBrowser, and by analogy HiBrowse, is therefore, at least in principle, no more complicated than using Galaxy in general. Another highly important reason for using a Galaxy-based system is the possibility of sharing results through the history sharing and Galaxy Pages functionality. This functionality makes it easy for other scientists to reproduce and scrutinize all results, by simply uploading the histories into their own Galaxy and re-run analyses, with the possibility of altering parameters. Such functionality is needed in computational research, as illustrated by the increased focus on reproducibility, and the realization that many publications in top journals do not contain enough methodological detail to be reproducible [225]. Another interesting feature of Galaxy is the ToolShed [226], where tools can be uploaded and immediately utilized through the GUI. In the future, it would be preferable to implement HiBrowse as a Galaxy tool that can be uploaded through the ToolShed. However, currently this is not possible, due to the limited track type support in Galaxy.

The alternative to using Galaxy, and a graphical user interface, would be a command-line based tool. Currently, only a relatively small set of tools are available for analysis of Hi-C (and similar) data, all of which are command-line based. In the program INSP3CT (mentioned in section 1.3.4.1), for example, basic 3D co-localization analysis of a list of genomic elements can be performed. In the HOMER tool (see section 1.3.4.2), the main Hi-C related features are data pre-processing, compartment analysis and detection of significant interactions, with some additional functionality for comparing two treatments. The advantage of using a command-line tool, instead of a GUI-based tool, is that it is easy to build pipelines and to extend software. Also, many bioinformaticians and software developers prefer to use command-line based tools instead of GUI-based tools. An attractive alternative is of course

to allow both command-line and GUI-based solutions. In the Genomic HyperBrowser, and in HiBrowse, analyses can be run using a command-line like language through the GUI, utilizing batch execution functionality. However, such functionality cannot completely replace classical command-line functionality.

One issue with the current GUI of the Genomic HyperBrowser, and therefore of HiBrowse, is that users are required, to some extent, to know beforehand the steps needed in order to perform a full analysis. For example, in order to do a hypothesis test of whether a set of linked elements are co-localized (in 3D), users usually have to first convert their data to a linked track, and only then is it possible to run the analysis on the resulting file. One of the reasons for doing it in this way is to build on the principles of Galaxy as a workflow-based system, where smaller analysis steps are combined into a workflow, ending up with a final result. This principle is similar to the principle of command-line based tools, but with the added benefit of having a graphical interface. One way of avoiding this conversion step is of course to have data available in linked (GTrack) format in the first place, but this will only be a realistic alternative if external upstream processing tools support the GTrack format. Similarly, uploading of Hi-C and other types of 3C-based data is currently possible in two ways. Either a GTrack file is uploaded directly (assuming that this was the format already used) or a specialized uploading-tool is used. In the uploading-tool, any tabulated format (including Excel) can be uploaded and then seamlessly converted into GTrack format. At the moment, however, this process can be time-consuming, since the system runs a check on all uploaded files to confirm that they are of valid format. Since data are represented as graphs in the system, this process requires looping over all edges to check if the file is correctly specified.

4.3 Biological relevance and usability

In **Paper I**, a set of topological properties underlying chromatin structural data was presented, together with a statistical test and an enrichment score for 3D co-localization. Due to the somewhat theoretical focus, the main audience for the tools presented is probably developers and advanced users looking to build upon the methods perhaps with the aim of tool implementation. We do also provide a basic implementation of the method in the Genomic HyperBrowser, allowing biologists and applied users to use some of the proposed tools. In **Paper II**, however, the set of tools are expanded and put into a more applicable context in the form of a web-tool for broad analysis of 3D co-localization and similar types of analyses of Hi-C data. The main audience for **Paper II** is therefore biologists and other applied researchers with a clear biological question in mind, seeking to analyze either own or publicly available data without relying on development, implementation and bug-testing on their own. The main target for **Paper III** are bioinformaticians and developers analyzing ChIA-PET data, and not directly aimed towards biologists without computational experience. While the paper also is accompanied by code for running analyses, this code is command-line based and presented mostly for reproducibility purposes. An aim for the future is to also implement these tools into HiBrowse.

The established tool for analysis of ChIA-PET data is the ChIA-PET tool ([173], discussed in sections 1.3.4.2 and 1.3.5.2). This is a command-line based tool, with relevant

software for performing a full analysis, from mapping of sequences to visualization of results. The method implemented in **Paper III** is not meant to be a direct competitor of this tool, even if it was to be implemented in the HiBrowse system (**Paper II**). The main aim of **Paper III** was to improve on the statistical methods and the underlying assumptions of the models for the step where significance of interactions is analyzed. Whether or not this method will become adapted in the community remains to be seen. Obviously, the usability of this method also depends on how much the ChIA-PET technology will be applied to new projects in the coming years.

It is clear that, at the moment, the tools presented in **Paper I** and **Paper II** are meant mostly for 3D co-localization and similar analysis at a larger scale. These types of analyses are mostly meant to elucidate spatial patterns such as transcription factories, proximities between a selected set of elements, contacts between insulator regions, and relationships between structural and functional features that govern the processes of the cell. Such types of analyses are in practice often part of a larger study of some sort, where one of the aspects is related to finding out whether structural features of the genome are relevant for the set of elements under study. Several examples from the literature illustrate the versatile and general nature of 3D co-localization analyses (see Table 4.2).

Table 4.2: Examples of 3D co-localization analyses where HiBrowse functionality could have been applied

Research question	**Query set**	**HiBrowse statistic**	**Refs**
Are origins of early replication co-localized in 3D?	Set with origins of replication	Co-localized in 3D?	[98, 99]
Are centromeres co-localized in 3D?	Set with centromeric positions	Co-localized in 3D?	[98, 99]
Are transcription factor (TF) target genes co-localized in 3D?	Gene sets with targets of various TFs	Linked elements co-localized in 3D?	[99, 189]
Are pairs of regions involved in translocation events co-localized in 3D?	Pairs of genomic positions involved in translocations	Linked elements co-localized in 3D? (possibly maintaining open/closed compartments)	[153]
Are evolutionary breakpoints between human and mouse co-localized in 3D?	Pairs of genomic positions involved in breakpoints	Linked elements co-localized in 3D? (possibly maintaining open/closed compartments)	[192]
Identify the differential 3D contacts between embryonic and cortex cell-lines in mouse	-	Identify significant differences between two 3D tracks	[88]
Identify the differential 3D contacts between cancerous and normal prostate cell-lines	-	Identify significant differences between two 3D tracks	[203]

The examples listed in Table 4.2 are meant to provide a few examples picked from the literature, and are not comprehensive. Nevertheless, it seems that functionality related to analysis of 3D co-localization, as implemented in HiBrowse, is definitively relevant. However,

what about the tools available in HiBrowse where no corresponding analysis is found in the published literature? Specifically, this relates to the statistics regarding the 3D co-localization between two point tracks, and the analysis of linked elements marked by "case" and "control" (see Figure 3.2). These statistical tests were implemented in HiBrowse in the hope that they could be of use in particular cases. For example, it is possible to imagine that it could be relevant to analyze the 3D co-localization between transcription factor (TF) binding sites and a set of genes, without specifying any information about any particular interactions between selected TF binding sites and genes. Such an analysis could be performed easily with the "Track 1 more/less co-localized with track 2 in 3D?" statistic. Additionally, the "Case-links more/less co-localized in 3D?" statistic could be used in settings where a subset of features picked from a larger set of interactions is studied. For example, if a researcher wants to know if a set of genes from a particular pathway taken from the KEGG database [227] is more co-localized in 3D than genes within KEGG pathways in general, this statistic could be used.

While it is clearly interesting to analyze whether a set of genes is more proximal within the nucleus, another point to consider is whether or not it is relevant to ask if a set of genes is *less* co-localized in 3D, than what could be expected by chance. This analysis option, in addition to the option of testing for a *difference* in 3D co-localization, was introduced in **Paper I** and **Paper II** for completeness only. It is difficult to imagine where such an analysis will be relevant, but it is also important not to exclude such cases. Perhaps analysis relating to the positioning at the peripheral parts of the nucleus, such as genomic regions associated with the nuclear lamina, will be a possible use case of such an analysis?

Even though much of the focus of this project has been on hypothesis testing and inference, some effort has been put into devising methods for quantification of 3D co-localization as well. In **Paper I**, we developed an enrichment score giving the degree of 3D proximity as compared to the expected background signal given by genomic distance and other factors that may be relevant to take into account. We also specifically stated in **Paper I** that the enrichment score and statistical significance should both be part of a given 3D co-localization analysis. This is an important point, considering that the ease of reaching significance increases with the size of the dataset. It is, however, inherently difficult to know what constitutes a biologically meaningful contact enrichment. Our example in **Paper I**, concerning the 3D proximity of mutational events in leukemia, illustrates this point. In the paper, we identified a rather strikingly significant 3D co-localization between regions involved with somatic mutations for intra-chromosomal interactions, when comparing across different resolutions, and when correcting for properties such as compartments and relative positioning along chromosome arms. However, as we pointed out in the paper, the enrichment scores were in the range 0.14-2.43%, and were too low to draw definite conclusions regarding biological relevance. How large the enrichment score needs to be before a biologically relevant finding is seen, however, is an open question. The best solution is probably to let the scientific community work out the relevant thresholds over time, based on accumulating experience. In this regard, it is important to note that in FISH studies of 3D co-localization, the enrichment of co-localization as measured for example by the number of times a pair of loci co-localize, has been found to be quite low. For example, in a paper where both 4C and cryo-FISH was applied to selected regions, significantly interacting regions were often found to co-localize at frequencies as low as a few percent [188]. This is expected, considering the great variability

of genome architecture across cells (see section 1.1.3.7).

We have also focused on investigating the effect of using different resolutions, and on analyzing inter- and intrachromosomal interactions, both separately and jointly. We pointed out in **Paper I** that analyses of 3D co-localization should preferentially be performed on different resolutions, since both P-values and enrichment scores can vary depending on resolution. We also pointed out the possibility of performing analyses on intra- and interchromosomal regions, both separately and jointly, since again, results and interpretations can be different for these two types of interactions. Considering this, it is important to note that running analyses on many different resolutions to fish for statistical significance is not a valid option. Since the bin-size used is dependent on the choice of the investigator, one may end up in a situation similar to what is known as the "modifiable areal unit problem" in geography [228], where aggregation of data into somewhat arbitrary groups (such as districts or postcode areas) can lead to a selection bias in the reported patterns. We have in **Paper I** proposed to report results on several bin-size resolutions, and proposed to investigate whether the P-values stabilize at a particular value. In any case, obtaining similar results at several different resolutions could be a minimal requirement for reporting a result. This approach has, to some extent, been established when analyzing 4C-data [167]. In the HiBrowse system (**Paper II**), we have facilitated the use of different bin-sizes by providing most datasets at four different resolutions (see Table 4.1). It is, however, up to the user to decide the appropriate resolution for a given analysis, and whether or not it is necessary to perform analyses on multiple resolutions.

Another point that was investigated in **Paper I**, was the *specificity* of 3D co-localization, when considering a given Hi-C dataset. This is important to consider since proximity, in the form of 3D co-localization between a selected set of regions, does not necessarily mean that this set is specifically the set that allows such 3D co-localization to occur. For example, since chromatin exists in the nucleus as a polymer chain, nearby elements will necessarily share spatial characteristics, but may actually be even more proximal than the initially selected set. Additionally, the positions selected might be found in regions where general 3D co-localization of elements is the norm, rather than a special case. To investigate this, we picked a query set of elements that had recently been shown to form clusters of co-localized elements ([130], explored in detail in **Paper III**). We calculated the enrichment of 3D co-localization and the P-value for this set, and then permuted the elements by iteratively shifting them one bin in a random direction. By re-calculating P-values and enrichment scores at each step, we were able to plot a curve showing how specific such 3D co-localization appeared to be. We found that after shifting 2-3 bins (corresponding to 200-300 kilobases), 3D co-localization enrichment was low and no longer significant. We did the same with the set of regions containing somatic mutations, and found similar results, albeit with much lower enrichment scores.

A large part of **Paper I** concerns the need for taking well-known topological properties of chromatin into account for valid hypothesis testing and for obtaining biologically realistic results. It is, however, often difficult to decide which exact properties that should be taken into account for a given analysis. While it is easy to see that direct physical properties of chromatin, such as genomic distance and transitivity relations need to be considered, properties such as domain architecture and relative positioning along chromosome arms may not be immediately obvious to correct for. If for example, a user wishes to analyze whether a set

of specific genes are proximal, taking into account a property such as compartmental structure (A and B compartments) should probably only be considered if the user *a priori* knows that the genes are positioned in a biased manner with regards to compartments. Since it is known that A-compartments are more gene-rich, and that spatial proximity is higher within compartments than between, comparing a set of genes biased towards A-compartments will almost always result in significance (as is shown in **Paper I**). In such a setting, not taking into account genomic compartments is probably similar to testing whether there is a bias towards one of the compartments, an analysis which can be done in a much simpler way than through 3D co-localization analysis. On the other hand, if the investigator has no prior expectation on compartmental positioning of the elements in the query set, it might be important to consider that compartmentalization could be the result of clustering of active regions in the first place (and not the other way around). In such cases, it may not be relevant to take into account compartmental structure. Therefore, in **Paper I** and **Paper II**, we have focused on allowing for the possibility of correcting for such factors, but leaving it up to the individual investigator to decide in practice whether compartments should be taken into account.

We are not the first to mention taking general properties, such as open/closed chromatin, into account. For example, Véron et al. [192] argued for taking both DNase sensitivity and gene density into account when analyzing 3D proximity of evolutionary breakpoints in the human genome. In their permutation-based test, they did this by dividing the genome into classes (open and closed) based on the DNase sensitivity. They argued that since breakpoints were much more likely to be positioned in open and gene-rich parts of the genome, it was necessary to take these properties into account during the permutations. A similar approach, correcting for chromatin compartments, was recently taken by Engreitz et al. [153].

At the moment, the most common types of analyses for Hi-C and similar 3C-based data are based on visualization and explorative analysis (see section 1.3.5.2). In HiBrowse (**Paper II**), we also have developed some tools for visual exploration of results. In Figure 4.1, two examples of Hi-C data visualizations, as produced by HiBrowse, are shown. Two main approaches are used in these visualizations. By representing the interaction frequencies as a heat map and clustering this based on hierarchical clustering, clusters of increased interaction frequencies between sets of elements can be identified. On the other hand, HiBrowse also allows representing the interaction frequencies between a selected set of elements as a graph, and spatially aligning nodes such that the lengths of their edges depend on the interaction frequencies. Analyses based on visualization of such data are challenging, however, since it is not immediately obvious how to interpret the resulting figures. Currently, the resulting graphics in HiBrowse is limited to static images, such as the ones shown in Figure 4.1. It is likely that interactive graphics, where users can reposition and select elements in the figure, to reveal further details on the underlying analysis, will be a more usable representation of these results. An aim for the future development of HiBrowse is therefore to allow such interactive graphical results. There may also be better ways of representing these data than using the classical heat map or graph representations. For example, as mentioned in section 1.3.5.2, some tools allowing for direct visualization of 3D structures have already been developed. It is unclear, however, whether such direct visualization will drastically improve our understanding of the underlying data. The challenge of visualizations like these is always to be able to balance complexity and interpretability. More work will therefore be needed to

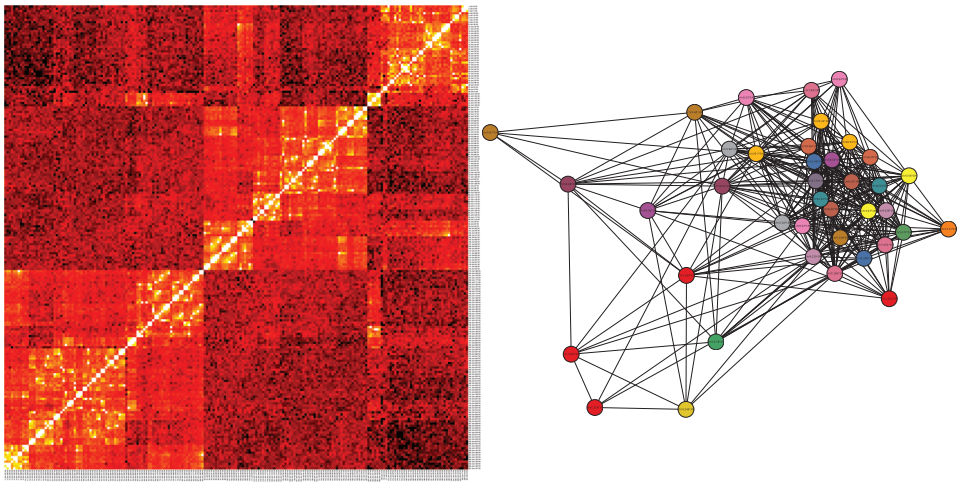

Figure 4.1: Examples of visualization of results using HiBrowse (**Paper II**). Left: Heat map visualization of clustered Hi-C data of all interaction frequencies on chromosome 1 in the GM06990 cell line. Right: Network visualization of Hi-C interactions between genes relevant for embryonic development (defined according to the GO category 'embryonic development') for the hESC cell line.

find optimal ways to visualize these data.

Additionally, further development of methods for identification of underlying structures based on Hi-C (and similar) data will be needed. It is important to point out that certain spatial arrangements of genomic elements will only be detected by considering the underlying 3D structure. For example, the radial positioning of elements relative to the rest of the chromosomes, such as in peripheral parts of the nucleus, is not necessarily detected by only considering 3D co-localization of elements. It is, however, a major challenge for 3D structure prediction based on Hi-C maps that the underlying data is based on a collection of a large number of cells. With further development and application of single-cell Hi-C technology (see section 1.2.6), structural modeling is likely to be much more robust and usable. However, even so, statistical analysis of 3D co-localization will still be relevant, since simply visualizing structural models of chromatin will not suffice to answer all questions regarding the spatial arrangements of genomic loci in such structures.

4.4 Future perspectives

Since techniques such as Hi-C and ChIA-PET are so novel, a natural question to ask is whether these technologies soon will be replaced by new technologies with completely different properties. While it is likely that these methods are improved and even replaced by better methods in the future, it is important to note that the topics touched upon in this thesis are likely to be relevant regardless of technological approach, since basic properties such as topology and genomic distance are a consequence of the physical structure of chromatin, and not the technology itself. Other topics, such as bias-correction and pre-processing of data are more likely to be superfluous or at least replaced as technology improves.

It is clear that, even if the technologies themselves do not change substantially in the

immediate future, increased throughput due to enhancements in next-generation sequencing technologies will increase the resolution of 3C-based datasets. In practice, this will mean Hi-C datasets with higher resolutions and smaller bin-sizes. However, it is important to consider that doubling the resolution requires quadrupling the amount of sequencing data, due to the quadratic nature of the resulting Hi-C matrices.

As higher resolution of Hi-C data is obtained, identification of significant interactions will become more relevant, since the genomic regions involved in such interactions will be more specific. The goal of such analyses it often to identify regulatory targets of genes, to gain an understanding of regulatory mechanisms in specific cell types. Analysis of such mechanisms, for example promoter-enhancer interactions, is typically not possible using 3D co-localization analysis, both due to the current resolution of Hi-C data, and due to the fact that such interactions are usually considered in an individual fashion. Even though one of the analyses in **Paper I** concerns promoters and enhancers, this analysis was meant to illustrate the power of the method, and to illustrate that active parts of the genome are closer together in 3D, even after correcting for compartmental structure. The methods presented in **Paper III**, however, are specifically geared towards the identification of cis-regulatory contacts, but using ChIA-PET data instead of Hi-C data. Despite the much lower resolution of Hi-C data, it has recently been shown that some of the more recent Hi-C data sets can be used for identification of such regulatory interactions as well [196] (see section 1.3.4.2). For this reason, a goal for the future is to implement statistics to identify significant interactions for Hi-C data and make them available in HiBrowse. It is likely that some of the insights gained in **Paper III** can be used to develop statistical tools for identification of significant interactions in Hi-C datasets as well. Likewise, it is also likely that analysis of 3D co-localization will be more usable as resolution increases. Another relevant consideration is whether it would be possible to analyze 3D co-localization for ChIA-PET data in a similar way as done for Hi-C data. This would, in principle, be analyzable using the tools developed for linked query sets (see Figure 3.2). However, at the moment, the resulting contact matrix from ChIA-PET experiments may be too sparse for such analyses to make sense.

The increased resolution and size of 3C-based datasets also signifies a need for re-thinking the way such data are represented. Currently, no standardized way of representing for example Hi-C data has been established. For example, plain text files containing symmetric matrices of all-versus-all chromosomes, configured BED files, or other types of in-house formats have been used. With the GTrack format [219], providing a unified way of allowing links between elements, a standardized way of representing such data was proposed. However, in order to store, retrieve and analyze large Hi-C, ChIA-PET, and other large-scale 3C-based datasets efficiently, novel ways of representing such data are likely to be needed.

Due to the establishment of large consortia such as ENCODE [4], Epigenome Roadmap [5] and FANTOM5 [229], the positions of cell-type specific regulatory elements are mapped in extreme detail. The challenge in the future, however, is to link these elements together and obtain a functional understanding of the orchestration of the regulatory program of the cell-types. Both Hi-C and ChIA-PET data are likely to be extremely important to map such regulatory interactions. The statistical tools presented in **Paper III** can be relevant for obtaining high-confidence interactions between regulatory elements. To do so, however, requires that many of the cell-lines already investigated are additionally analyzed with methods such

as Hi-C and ChIA-PET.

With ever-increasing amounts of epigenetic data, it is clear that approaches for integrative analysis are needed to fully map the vast amounts of knowledge hidden in the combinatory patterns of each individual data track. A very illustrative example of the power of integrative methods was given in section 1.3.4.4, where combinations of epigenetic signals across different cell-lines were shown to allow for identification of interactions between distal regions. Other potentially powerful approaches are likely to emerge when considering that diverse epigenetic datasets can be combined and integrated in a myriad of ways. Some tools aiming at such combinatorial approaches were mentioned in section 1.3.6, but other tools and resources are likely to be needed in addition to these.

It is probable that the demand for tools with built-in functionality for reproducible analysis will continue to rise [225]. The HiBrowse tool developed in **Paper II** provides such functionality through allowing results, together with all the settings and parameters, to be inspected and re-run through the Galaxy Pages functionality. In this way, results can also easily be shared between researchers working on the same project, again functionality likely to be more important in the future.

Interestingly, several novel uses of genome-wide chromosome conformation capture data have emerged in the last years. For example, a promising new application in genome assembly exploits the fact that interaction frequencies are more frequent for regions in close genomic distance on the same chromosome [230]. This approach has in fact been found to be able to close gaps between contigs that could not be solved using conventional sequencing methods [231]. Also, in metagenomics studies, where the aim is to identify and characterize the abundance of microbial communities in a given sample, the same property has proven useful. Since a major challenge of such studies is to identify the clusters of DNA sequences coming from the same species, it is of major help to aid such clustering by using information of physically proximal sequences [232]. Additionally, but perhaps much more far-fetched, Hi-C and similar proximity-ligation based techniques have even been proposed to be able to aid in the mapping of connections in the brain [233].

While it is difficult to predict what will happen in the future, it is clear that taking the three-dimensional perspective on the nuclear organization of chromatin is a step in the right direction for understanding how cellular function is regulated and orchestrated. Methods for analysis and processing of such datasets, built on robust methodology and valid assumptions, will therefore be needed and further developed in the years to come.

Chapter 5

Conclusions

While data relating to the positioning of elements along the one-dimensional genome are mapped at an ever-increasing pace, information regarding how these elements are positioned, organized and regulated in three dimensions is needed to obtain a realistic understanding of the complexity underlying the orchestration of the regulatory mechanisms of the genome. Importantly, unraveling this complexity is likely to be vital for understanding and preventing human genetic diseases, since much of the identified genetic variation relevant to human disease has been found to reside outside of genes. Recently, it has been speculated that much of this variation affects regulatory mechanisms involving physical interactions between distal regulatory sites, and identification of the target genes is therefore highly relevant for development of therapeutic drugs.

Recent technologies coupling chromosome conformation capture to next-generation sequencing, such as Hi-C and ChIA-PET, allow for genome-wide identification of 3D contacts at unprecedented resolution. However, analyzing such data is challenging due to the complexity of the underlying chromatin structure. During the PhD-project presented in this thesis, some statistical methods for analysis of Hi-C and related data have been presented, focusing on taking into account important properties inherent in such data. The tools have been implemented in a web-based software suit called HiBrowse, where users can analyze various aspects relating to the 3D structure of chromatin, either by uploading their own data, or by basing analysis on publicly available data. A statistical method for analysis of ChIA-PET data has also been presented, allowing for robust inferential analysis of genome-wide regulatory interactions.

Taken together, the results in this thesis point to the importance of applying realistic assumptions when performing inferential analysis on chromosome conformation capture data. With the rapid increase of genome-wide mapping of epigenomic datasets across tissues and cell-lines, it is likely that the methods and tools presented in this thesis can provide a well grounded methodological framework for further studies into the three-dimensional organization of genomes.

References

1. Darwin C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. New York: D. Appleton 1859.
2. Henig RM. The monk in the garden: the lost and found genius of Gregor Mendel, the father of genetics. Houghton Mifflin Harcourt, 2000.
3. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
4. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
5. Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH roadmap epigenomics mapping consortium. Nature biotechnology 2010;28:1045–1048.
6. Annunziato A. DNA packaging: Nucleosomes and chromatin. Nature Education 2008;1:26.
7. Haeckel EHPA. Generelle Morphologie der Organismen: allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von Charles Darwin reformirte Descendenz-Theorie. Vol. 2. G. Reimer, 1866.
8. Miescher-Rüsch F. Ueber die chemische Zusammensetzung der Eiterzellen. 1871.
9. Flemming W. Zellsubstanz, kern und zelltheilung. Vogel, 1882.
10. Van Beneden E. Recherches sur la maturation de l'oeuf, la fécondation, et la division cellulaire. H. Engelcke, 1883.
11. O H. Jenaische Z. Naturwiss 1885;18:276–318.
12. Boveri T. Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns. Verhandl. d. Phys.-med. Gesellsch. Würzburg. NF Bd 1902;35.
13. Sutton WS. The chromosomes in heredity. The Biological Bulletin 1903;4:231–250.
14. Avery OT, MacLeod CM, and McCarty M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. The Journal of experimental medicine 1944;79:137–158.
15. Hershey AD and Chase M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. The Journal of general physiology 1952;36:39–56.
16. Watson JD and Crick F. Molecular structure of nucleic acids. Nature 1953;171:737–738.

17. Crick FH. On protein synthesis. In: *Symposia of the Society for Experimental Biology*. Vol. 12. 1958:138.

18. Allfrey V, Faulkner R, and Mirsky A. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. Proceedings of the National Academy of Sciences 1964;51:786.

19. Pauling L, Corey RB, and Branson HR. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proceedings of the National Academy of Sciences 1951;37:205–211.

20. Pauling L and Corey RB. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proceedings of the National Academy of Sciences 1951;37:729.

21. Franklin RE and Gosling RG. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. Nature 1953;172:156–157.

22. Pardon J, Wilkins M, and Richards B. Super-helical model for nucleohistone. Nature 1967;215:508–509.

23. Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science 1974;184:868–871.

24. Olins DE and Olins AL. Chromatin history: our view from the bridge. Nature reviews Molecular cell biology 2003;4:809–814.

25. Finch J and Klug A. Solenoidal model for superstructure in chromatin. Proceedings of the National Academy of Sciences 1976;73:1897–1901.

26. Rappold GA, Cremer T, Hager H, Davies K, Müller C, and Yang T. Sex chromosome positions in human interphase nuclei as studied by in situ hybridization with chromosome specific DNA probes. Human genetics 1984;67:317–325.

27. Comings DE. The rationale for an ordered arrangement of chromatin in the interphase nucleus. American journal of human genetics 1968;20:440.

28. Lewin R. Do Chromosomes Cross Talk? Science 1981;214:1334–1335.

29. Cremer T and Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature reviews genetics 2001;2:292–301.

30. Maeshima K, Hihara S, and Eltsov M. Chromatin structure: does the 30-nm fibre exist in vivo? Current opinion in cell biology 2010;22:291–297.

31. Nishino Y, Eltsov M, Joti Y, et al. Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. The EMBO journal 2012;31:1644–1653.

32. Novo CL and Londoño-Vallejo JA. Telomeres and the nucleus. In: *Seminars in cancer biology*. Vol. 23. 2. Elsevier. 2013:116–124.

33. Heitz E. Heterochromatin, chromocentren, chromomeren. Komm. Fischer, 1929.

34. Schultz J. The function of heterochromatin. In: *Proc. 7th Int. Genet. Congr*. 1939:257–62.

35. Latchman DS and McClements M. Gene control. Garland Science, 2010.

36. Caspersson T, Farber S, Foley G, et al. Chemical differentiation along metaphase chromosomes. Experimental cell research 1968;49:219–222.

37. Fatemi M, Pao MM, Jeong S, et al. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic acids research 2005;33:e176–e176.

38. Schones DE and Zhao K. Genome-wide approaches to studying chromatin modifications. Nature Reviews Genetics 2008;9:179–191.

39. Ozsolak F, Song JS, Liu XS, and Fisher DE. High-throughput mapping of the chromatin structure of human promoters. Nature biotechnology 2007;25:244–248.

40. Ong CT and Corces VG. CTCF: an architectural protein bridging genome topology and function. Nature Reviews Genetics 2014.

41. Hon G, Wang W, and Ren B. Discovery and annotation of functional chromatin signatures in the human genome. PLoS computational biology 2009;5:e1000566.

42. Ernst J and Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology 2010;28:817–825.

43. Filion GJ, Bemmel JG van, Braunschweig U, et al. Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. Cell 2010;143:212–224.

44. Calo E and Wysocka J. Modification of enhancer chromatin: what, how, and why? Molecular cell 2013;49:825–837.

45. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473:43–49.

46. Kharchenko PV, Alekseyenko AA, Schwartz YB, et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 2011;471:480–485.

47. Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, et al. Extensive variation in chromatin states across humans. Science 2013;342:750–752.

48. Bickmore WA and Steensel B van. Genome architecture: domain organization of interphase chromosomes. Cell 2013;152:1270–1284.

49. McBryant SJ, Adams VH, and Hansen JC. Chromatin architectural proteins. Chromosome Research 2006;14:39–51.

50. Taverna SD, Li H, Ruthenburg AJ, Allis CD, and Patel DJ. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. Nature structural & molecular biology 2007;14:1025–1040.

51. Magistri M, Faghihi MA, St Laurent III G, and Wahlestedt C. Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. Trends in genetics 2012;28:389–396.

52. Kung JT and Lee JT. RNA in the Loop. Developmental cell 2013;24:565–567.

53. Ørom UA and Shiekhattar R. Long non-coding RNAs and enhancers. Current opinion in genetics & development 2011;21:194–198.

54. Nagano T and Fraser P. No-nonsense functions for long noncoding RNAs. Cell 2011;145:178–181.

55. Ohno S, Kaplan W, and Kinosita R. Formation of the sex chromatin by a single X-chromosome in liver cells of Rattus norvegicus. Experimental cell research 1959;18:415–418.

56. Dekker J, Rippe K, Dekker M, and Kleckner N. Capturing chromosome conformation. Science 2002;295:1306–1311.

57. Croft JA, Bridger JM, Boyle S, Perry P, Teague P, and Bickmore WA. Differences in the localization and morphology of chromosomes in the human nucleus. The Journal of cell biology 1999;145:1119–1131.

58. Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis JA, and Bickmore WA. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Human molecular genetics 2001;10:211–219.

59. Saccone S, Federico C, and Bernardi G. Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. Gene 2002;300:169–178.

60. Chambeyron S and Bickmore WA. Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. Genes & development 2004;18:1119–1130.

61. Ragoczy T, Bender M, Telling A, Byron R, and Groudine M. The locus control region is required for association of the murine β-globin locus with engaged transcription factories during erythroid maturation. Genes & development 2006;20:1447–1457.

62. Takizawa T, Meaburn KJ, and Misteli T. The meaning of gene positioning. Cell 2008;135:9–13.

63. Jackson D and Cook P. Transcription occurs at a nucleoskeleton. The EMBO journal 1985;4:919.

64. Jackson DA, Hassan AB, Errington RJ, and Cook PR. Visualization of focal sites of transcription within human nuclei. The EMBO journal 1993;12:1059.

65. Iborra FJ, Pombo A, Mcmanus J, Jackson DA, and Cook PR. The topology of transcription by immobilized polymerases. Experimental cell research 1996;229:167–173.

66. Osborne CS, Chakalova L, Brown KE, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. Nature genetics 2004;36:1065–1071.

67. Sutherland H and Bickmore WA. Transcription factories: gene expression in unions? Nature Reviews Genetics 2009;10:457–466.

68. Martin S and Pombo A. Transcription factories: quantitative studies of nanostructures in the mammalian nucleus. Chromosome Research 2003;11:461–470.

69. Mitchell JA and Fraser P. Transcription factories are nuclear subcompartments that remain in the absence of transcription. Genes & development 2008;22:20–25.

70. Chepelev I, Wei G, Wangsa D, Tang Q, and Zhao K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell research 2012;22:490–503.

71. Di Stefano M, Rosa A, Belcastro V, Bernardo D di, and Micheletti C. Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome 19. PLoS computational biology 2013;9:e1003019.

72. Schoenfelder S, Sexton T, Chakalova L, et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nature genetics 2010;42:53–61.

73. Pederson T. The nucleolus. Cold Spring Harbor perspectives in biology 2011;3:a000638.

74. Dillon N. Gene regulation and large-scale chromatin organization in the nucleus. Chromosome Research 2006;14:117–126.

75. Németh A, Conesa A, Santoyo-Lopez J, et al. Initial genomics of the human nucleolus. PLoS genetics 2010;6:e1000889.

76. Koningsbruggen S van, Gierliński M, Schofield P, et al. High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. Molecular biology of the cell 2010;21:3735–3748.

77. Gerace L, Blum A, and Blobel G. Immunocytochemical localization of the major polypeptides of the nuclear pore complex-lamina fraction. Interphase and mitotic distribution. The Journal of cell biology 1978;79:546–566.

78. Fisher P. Chromosomes and chromatin structure: the extrachromosomal karyoskeleton. Current opinion in cell biology 1989;1:447–453.

79. Burke B and Stewart CL. The nuclear lamins: flexibility in function. Nature Reviews Molecular Cell Biology 2013;14:13–24.

80. Shevelyov YY and Nurminsky DI. The nuclear lamina as a gene-silencing hub. Current issues in molecular biology 2012;14:27.

81. Finlan LE, Sproul D, Thomson I, et al. Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS genetics 2008;4:e1000039.

82. Reddy K, Zullo J, Bertolino E, and Singh H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. Nature 2008;452:243–247.

83. Kumaran RI and Spector DL. A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. The Journal of cell biology 2008;180:51–65.

84. Guelen L, Pagie L, Brasset E, et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 2008;453:948–951.

85. Lieberman-Aiden E, Berkum NL van, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326:289–293.

86. Imakaev M, Fudenberg G, McCord RP, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature methods 2012;9:999–1003.

87. Yaffe E and Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature genetics 2011;43:1059–1065.

88. Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485:376–380.

89. Phillips-Cremins JE, Sauria ME, Sanyal A, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell 2013;153:1281–1295.

90. Nora EP, Lajoie BR, Schulz EG, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 2012;485:381–385.

91. Dekker J, Marti-Renom MA, and Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics 2013.

92. Stambrook PJ and Flickinger R. Changes in chromosomal DNA replication patterns in developing frog embryos. Journal of Experimental Zoology 1970;174:101–113.

93. Heun P, Laroche T, Raghuraman M, and Gasser SM. The positioning and dynamics of origins of replication in the budding yeast nucleus. The Journal of cell biology 2001;152:385–400.

94. Yurov YB and Liapunova NA. The units of DNA replication in the mammalian chromosomes: evidence for a large size of replication units. Chromosoma 1977;60:253–267.

95. Hansen RS, Thomas S, Sandstrom R, et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proceedings of the National Academy of Sciences 2010;107:139–144.

96. Ryba T, Hiratani I, Lu J, et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome research 2010;20:761–770.

97. Yaffe E, Farkash-Amar S, Polten A, Yakhini Z, Tanay A, and Simon I. Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture. PLoS genetics 2010;6:e1001011.

98. Duan Z, Andronescu M, Schutz K, et al. A three-dimensional model of the yeast genome. Nature 2010;465:363–367.

99. Witten DM and Noble WS. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. Nucleic acids research 2012;40:3849–3855.

100. Nora EP, Dekker J, and Heard E. Segmental folding of chromosomes: A basis for structural and regulatory chromosomal neighborhoods? Bioessays 2013;35:818–828.

101. Pauler FM, Sloane MA, Huang R, et al. H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. Genome research 2009;19:221–233.

102. Wen B, Wu H, Shinkai Y, Irizarry RA, and Feinberg AP. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. Nature genetics 2009;41:246–250.

103. Kim TH, Abdullaev ZK, Smith AD, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 2007;128:1231–1245.

104. Jothi R, Cuddapah S, Barski A, Cui K, and Zhao K. Genome-wide identification of in vivo protein–DNA binding sites from ChIP-Seq data. Nucleic acids research 2008;36:5221–5231.

105. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, and Zhao K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome research 2009;19:24–32.

106. Gaszner M and Felsenfeld G. Insulators: exploiting transcriptional and epigenetic mechanisms. Nature Reviews Genetics 2006;7:703–713.

107. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, and Lander ES. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proceedings of the National Academy of Sciences 2007;104:7145–7150.

108. Botta M, Haider S, Leung IX, Lio P, and Mozziconacci J. Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. Molecular systems biology 2010;6.

109. Phillips JE and Corces VG. CTCF: master weaver of the genome. Cell 2009;137:1194–1211.

110. Handoko L, Xu H, Li G, et al. CTCF-mediated functional chromatin interactome in pluripotent cells. Nature genetics 2011;43:630–638.

111. Parelho V, Hadjur S, Spivakov M, et al. Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell 2008;132:422–433.

112. Rubio ED, Reiss DJ, Welcsh PL, et al. CTCF physically links cohesin to chromatin. Proceedings of the National Academy of Sciences 2008;105:8309–8314.

113. Wendt KS, Yoshida K, Itoh T, et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature 2008;451:796–801.

114. Wendt KS and Peters JM. How cohesin and CTCF cooperate in regulating gene expression. Chromosome research 2009;17:201–214.

115. Zuin J, Dixon JR, Reijden MI van der, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proceedings of the National Academy of Sciences 2014;111:996–1001.

116. Cavalli G and Misteli T. Functional implications of genome topology. Nature structural & molecular biology 2013;20:290–299.

117. Blackwood EM and Kadonaga JT. Going the distance: a current view of enhancer action. Science 1998;281:60–63.

118. Bulger M and Groudine M. Looping versus linking: toward a model for long-distance gene activation. Genes & development 1999;13:2465–2477.

119. Hou C and Corces VG. Throwing transcription for a loop: expression of the genome in the 3D nucleus. Chromosoma 2012;121:107–116.

120. Sanyal A, Lajoie BR, Jain G, and Dekker J. The long-range interaction landscape of gene promoters. Nature 2012;489:109–113.

121. Laat W de and Grosveld F. Spatial organization of gene expression: the active chromatin hub. Chromosome Research 2003;11:447–459.

122. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, and Laat W de. Looping and interaction between hypersensitive sites in the active β-globin locus. Molecular cell 2002;10:1453–1465.

123. Fraser P, Pruzina S, Antoniou M, and Grosveld F. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. Genes & development 1993;7:106–113.

124. Palstra RJ, Tolhuis B, Splinter E, Nijmeijer R, Grosveld F, and Laat W de. The β-globin nuclear compartment in development and erythroid differentiation. Nature genetics 2003;35:190–194.

125. Li Q, Peterson KR, Fang X, and Stamatoyannopoulos G. Locus control regions. Blood 2002;100:3077–3086.

126. Spilianakis CG, Lalioti MD, Town T, Lee GR, and Flavell RA. Interchromosomal associations between alternatively expressed loci. Nature 2005;435:637–645.

127. Ling JQ, Li T, Hu JF, et al. CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. Science 2006;312:269–272.

128. Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, and Axel R. Interchromosomal interactions and olfactory receptor choice. Cell 2006;126:403–413.

129. Apostolou E and Thanos D. Virus Infection Induces NF-κB-Dependent Interchromosomal Associations Mediating Monoallelic IFN-β Gene Expression. Cell 2008;134:85–96.

130. Li G, Ruan X, Auerbach RK, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 2012;148:84–98.

131. Lanzuolo C, Roure V, Dekker J, Bantignies F, and Orlando V. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. Nature cell biology 2007;9:1167–1174.

132. Bantignies F, Roure V, Comet I, et al. Polycomb-Dependent Regulatory Contacts between Distant Hox Loci in Drosophila. Cell 2011;144:214–226.

133. Tiwari VK, Cope L, McGarvey KM, Ohm JE, and Baylin SB. A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations. Genome research 2008;18:1171–1179.

134. Tiwari VK, McGarvey KM, Licchesi JD, et al. PcG proteins, DNA methylation, and gene repression by chromatin looping. PLoS biology 2008;6:e306.

135. O'Sullivan JM, Tan-Wong SM, Morillon A, et al. Gene loops juxtapose promoters and terminators in yeast. Nature genetics 2004;36:1014–1018.

136. Ansari A and Hampsey M. A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping. Genes & development 2005;19:2969–2978.

137. Mapendano CK, Lykke-Andersen S, Kjems J, Bertrand E, and Jensen TH. Crosstalk between mRNA 3' end processing and transcription initiation. Molecular cell 2010;40:410–422.

138. Tan-Wong SM, French JD, Proudfoot NJ, and Brown MA. Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene. Proceedings of the National Academy of Sciences 2008;105:5160–5165.

139. O'Reilly D and Greaves DR. Cell-type-specific expression of the human CD68 gene is associated with changes in Pol II phosphorylation and short-range intrachromosomal gene looping. Genomics 2007;90:407–415.

140. Perkins KJ, Lusic M, Mitar I, Giacca M, and Proudfoot NJ. Transcription-dependent gene looping of the HIV-1 provirus is dictated by recognition of pre-mRNA processing signals. Molecular cell 2008;29:56–68.

141. Lanctôt C, Cheutin T, Cremer M, Cavalli G, and Cremer T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nature Reviews Genetics 2007;8:104–115.

142. Rapkin LM, Anchel DR, Li R, and Bazett-Jones DP. A view of the chromatin landscape. Micron 2012;43:150–158.

143. Kozubek S, Lukásová E, Jirsová P, et al. 3D structure of the human genome: order in randomness. Chromosoma 2002;111:321–331.

144. Chubb JR, Boyle S, Perry P, and Bickmore WA. Chromatin motion is constrained by association with nuclear compartments in human cells. Current Biology 2002;12:439–445.

145. Meaburn KJ, Gudla PR, Khan S, Lockett SJ, and Misteli T. Disease-specific gene repositioning in breast cancer. The Journal of cell biology 2009;187:801–812.

146. Marti-Renom MA and Mirny LA. Bridging the resolution gap in structural modeling of 3D genome organization. PLoS computational biology 2011;7:e1002125.

147. Baù D and Marti-Renom MA. Structure determination of genomic domains by satisfaction of spatial restraints. Chromosome research 2011;19:25–35.

148. Kalhor R, Tjong H, Jayathilaka N, Alber F, and Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nature biotechnology 2012;30:90–98.

149. Nagano T, Lubling Y, Stevens TJ, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 2013;502:59–64.

150. Naumova N, Imakaev M, Fudenberg G, et al. Organization of the mitotic chromosome. Science 2013;342:948–953.

151. Pomerantz MM, Ahmadiyeh N, Jia L, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nature genetics 2009;41:882–884.

152. Crutchley JL, Wang XQD, Ferraiuolo MA, and Dostie J. Chromatin conformation signatures: ideal human disease biomarkers? Biomarkers in medicine 2010;4:611–629.

153. Engreitz JM, Agarwala V, and Mirny LA. Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease. PloS one 2012;7:e44196.

154. De S and Michor F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nature biotechnology 2011;29:1103–1108.

155. Fudenberg G, Getz G, Meyerson M, and Mirny LA. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nature biotechnology 2011;29:1109–1113.

156. Schuster-Böckler B and Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 2012;488:504–507.

157. Worman HJ and Bonne G. “Laminopathies”: a wide spectrum of human diseases. Experimental cell research 2007;313:2121–2133.

158. Moindrot B, Bouvet P, and Mongelard F. Chromatin Structure and Organization: The Relation with Gene Expression During Development and Disease. Epigenetics: Development and Disease 2013:373–396.

159. Leung KN, Chamberlain SJ, Lalande M, and LaSalle JM. Neuronal chromatin dynamics of imprinting in development and disease. Journal of cellular biochemistry 2011;112:365–373.

160. Jakovcevski M and Akbarian S. Epigenetic mechanisms in neurological disease. Nature medicine 2012;18:1194–1204.

161. Van Steensel B and Dekker J. Genomics tools for unraveling chromosome architecture. Nature biotechnology 2010;28:1089–1095.

162. Volpi EV and Bridger JM. FISH glossary: an overview of the fluorescence in situ hybridization technique. Biotechniques 2008;45:385–386.

163. Bolzer A, Kreth G, Solovei I, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS biology 2005;3:e157.

164. Morey M, Fernández-Marmiesse A, Castiñeiras D, Fraga JM, Couce ML, and Cocho JA. A glimpse into past, present, and future DNA sequencing. Molecular genetics and metabolism 2013;110:3–24.

165. Wit E de and Laat W de. A decade of 3C technologies: insights into nuclear organization. Genes & development 2012;26:11–24.

166. Dekker J. The three'C's of chromosome conformation capture: controls, controls, controls. Nature Methods 2006;3:17–21.

167. Werken HJ van de, Landan G, Holwerda SJ, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nature methods 2012;9:969–972.

168. Simonis M, Kooren J, and De Laat W. An evaluation of 3C-based methods to capture DNA interactions. Nature methods 2007;4:895–901.

169. Rodley C, Bertels F, Jones B, and O'Sullivan J. Global identification of yeast chromosome interactions using genome conformation capture. Fungal Genetics and Biology 2009;46:879–886.

170. Dekker J and Mirny L. Biological techniques: Chromosomes captured one by one. Nature 2013;502:45–46.

171. Fullwood MJ, Liu MH, Pan YF, et al. An oestrogen-receptor-α-bound human chromatin interactome. Nature 2009;462:58–64.

172. Sajan SA and Hawkins RD. Methods for identifying higher-order chromatin structure. Annual review of genomics and human genetics 2012;13:59–82.

173. Li G, Fullwood MJ, Xu H, et al. Software ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biology 2010;11:R22.

174. Gascoigne DK, Taft RJ, Pheasant M, and Mattick JS. Reassessment of the Hi-C analysis of human genome architecture. 2011.

175. Hu M, Deng K, Qin Z, and Liu JS. Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. Quantitative Biology 2013;1:156–174.

176. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, and Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics 2012;28:3131–3133.

177. Fraser J, Rousseau M, Shenker S, et al. Chromatin conformation signatures of cellular differentiation. Genome biology 2009;10:R37.

178. Baù D and Marti-Renom MA. Genome structure determination via 3C-based data integration by the Integrative Modeling Platform. Methods 2012;58:300–306.

179. Baù D, Sanyal A, Lajoie BR, et al. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nature structural & molecular biology 2011;18:107–114.

180. Rosa A and Zimmer C. Computational models of large-scale genome architecture. International review of cell and molecular biology 2013;307:275–349.

181. Rousseau M, Fraser J, Ferraiuolo MA, Dostie J, and Blanchette M. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC bioinformatics 2011;12:414.

182. Hu M, Deng K, Qin Z, et al. Bayesian inference of spatial organizations of chromosomes. PLoS computational biology 2013;9:e1002893.

183. Fussner E, Ching RW, and Bazett-Jones DP. Living without 30nm chromatin fibers. Trends in biochemical sciences 2011;36:1–6.

184. Fudenberg G and Mirny LA. Higher-order chromatin structure: bridging physics and biology. Current opinion in genetics & development 2012;22:115–124.

185. Münkel C and Langowski J. Chromosome structure predicted by a polymer model. Physical Review E 1998;57:5888.

186. Mirny LA. The fractal globule as a model of chromatin architecture in the cell. Chromosome research 2011;19:37–51.

187. Rosa A and Everaers R. Structure and dynamics of interphase chromosomes. PLoS computational biology 2008;4:e1000153.

188. Simonis M, Klous P, Splinter E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nature genetics 2006;38:1348–1354.

189. Dai Z and Dai X. Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. Nucleic acids research 2012;40:27–36.

190. Ben-Elazar S, Yakhini Z, and Yanai I. Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. Nucleic acids research 2013;41:2191–2201.

191. Eden E, Lipson D, Yogev S, et al. Discovering motifs in ranked lists of DNA sequences. PLoS computational biology 2007;3:e39.

192. Véron AS, Lemaitre C, Gautier C, Lacroix V, and Sagot MF. Close 3D proximity of evolutionary breakpoints argues for the notion of spatial synteny. BMC genomics 2011;12:303.

193. Tanizawa H, Iwasaki O, Tanaka A, et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic acids research 2010;38:8164–8177.

194. Kruse K, Sewitz S, and Babu MM. A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. Nucleic acids research 2013;41:701–710.

195. Wang H, Duggal G, Patro R, Girvan M, Hannenhalli S, and Kingsford C. Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin. In: *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. ACM. 2013:306.

196. Ay F, Bailey TL, and Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome research 2014:doi:10.1101/gr.160374.113.

197. Lin YC, Benner C, Mansson R, et al. Global changes in the nuclear positioning of genes and intra-and interdomain genomic interactions that orchestrate B cell fate. Nature immunology 2012;13:1196–1204.

198. Benner C. HOMER. `http://homer.salk.edu/homer/`. Accessed: 05.03.2014.

199. Jin F, Li Y, Dixon JR, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 2013;503:290–294.

200. Williams RL, Starmer J, Mugford JW, et al. fourSig: a method for determining chromosomal interactions in 4C-Seq data. Nucleic acids research 2014:gku156.

201. Splinter E, Wit E de, Werken HJ van de, Klous P, and Laat W de. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. Methods 2012;58:221–230.

202. Thongjuea S, Stadhouders R, Grosveld FG, Soler E, and Lenhard B. r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. Nucleic acids research 2013;41:e132–e132.

203. Rickman DS, Soong TD, Moss B, et al. Oncogene-mediated alterations in chromatin conformation. Proceedings of the National Academy of Sciences 2012;109:9083–9088.

204. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. Nature 2012;489:75–82.

205. Shen Y, Yue F, McCleary DF, et al. A map of the cis-regulatory sequences in the mouse genome. Nature 2012;488:116–120.

206. Sexton T, Yaffe E, Kenigsberg E, et al. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell 2012;148:458–472.

207. Krzywinski M, Schein J, Birol İ, et al. Circos: an information aesthetic for comparative genomics. Genome research 2009;19:1639–1645.

208. Zhou X, Lowdon RF, Li D, et al. Exploring long-range genome interactions using the WashU Epigenome Browser. Nature methods 2013;10:375–376.

209. Servant N, Lajoie BR, Nora EP, et al. HiTC: exploration of high-throughput 'C'experiments. Bioinformatics 2012;28:2843–2844.

210. Li C, Dong X, Fan H, Wang C, Ding G, and Li Y. The 3DGD: a database of genome 3D structure. Bioinformatics 2014:btu081.

211. Asbury TM, Mitman M, Tang J, and Zheng WJ. Genome3D: A viewer-model framework for integrating and visualizing multi-scale epigenomic information within a three-dimensional genome. BMC bioinformatics 2010;11:444.

212. Shavit Y et al. CytoHiC: a cytoscape plugin for visual comparison of Hi-C networks. Bioinformatics 2013;29:1206–1207.

213. Li MJ, Wang LY, Xia Z, Sham PC, and Wang J. GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. Nucleic acids research 2013;41:W150–W158.

214. Font-Burgada J, Reina O, Rossell D, and Azorín F. chroGPS, a global chromatin positioning system for the functional analysis and visualization of the epigenome. Nucleic acids research 2013:gkt1186.

215. Sandve GK, Gundersen S, Rydbeck H, et al. The Genomic HyperBrowser: inferential genomics at the sequence level. Genome biology 2010;11:R121.

216. Sandve GK, Gundersen S, Johansen M, et al. The Genomic HyperBrowser: an analysis web server for genome-scale data. Nucleic acids research 2013;41:W133–W141.

217. Disanto G, Sandve GK, Berlanga-Taylor AJ, et al. Genomic regions associated with multiple sclerosis are active in B cells. PloS one 2012;7:e32281.

218. Disanto G, Sandve GK, Ricigliano VA, et al. DNase hypersensitive sites and association with multiple sclerosis. Human molecular genetics 2014;23:942–948.

219. Gundersen S, Kalaš M, Abul O, Frigessi A, Hovig E, and Sandve GK. Identifying elemental genomic track types and representing them uniformly. BMC bioinformatics 2011;12:494.

220. Robinson MD, McCarthy DJ, and Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139–140.

221. Goecks J, Nekrutenko A, Taylor J, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology 2010;11:R86.

222. Tukey JW. The future of data analysis. The Annals of Mathematical Statistics 1962:1–67.

223. Lemieux JE, Kyes SA, Otto TD, et al. Genome-wide profiling of chromosome interactions in Plasmodium falciparum characterizes nuclear architecture and reconfigurations associated with antigenic variation. Molecular microbiology 2013;90:519–537.

224. Moissiard G, Cokus SJ, Cary J, et al. MORC family ATPases required for heterochromatin condensation and gene silencing. Science 2012;336:1448–1451.

225. Sandve GK, Nekrutenko A, Taylor J, and Hovig E. Ten simple rules for reproducible computational research. PLoS computational biology 2013;9:e1003285.

226. Blankenberg D, Von Kuster G, Bouvier E, et al. Dissemination of scientific software with Galaxy ToolShed. Genome Biology 2014;15:403.

227. Kanehisa M. The KEGG database. 'In silico' simulation of biological processes 2002;247:91–103.

228. Openshaw S. The modifiable areal unit problem. Vol. 38. Geo Books Norwich, 1983.

229. The FANTOM Consortium. A promoter-level mammalian expression atlas. Nature 2014;507:462–470.

230. Korbel JO and Lee C. Genome assembly and haplotyping with Hi-C. Nature biotechnology 2013;31:1099–1101.

231. Kaplan N and Dekker J. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nature biotechnology 2013;31:1143.

232. Beitel CW, Froenicke L, Lang JM, et al. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. Tech. rep. PeerJ PrePrints, 2014.

233. Dekker J, Wysocka J, Mattaj I, Lieberman AE, and Pikaard C. Nuclear biology: what's been most surprising? Cell 2013;152:1207–1208.

I

5164–5174 *Nucleic Acids Research, 2013, Vol. 41, No. 10* *Published online 9 April 2013*
doi:10.1093/nar/gkt227

Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements

Jonas Paulsen[1], Tonje G. Lien[2], Geir Kjetil Sandve[3,4], Lars Holden[5], Ørnulf Borgan[2], Ingrid K. Glad[2] and Eivind Hovig[1,3,6,*]

[1]Section for Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, PO Box 4950, Nydalen, N-0424 Oslo, Norway, [2]Department of Mathematics, University of Oslo, PO Box 1053, Blindern, 0316 Oslo, Norway, [3]Department of Informatics, University of Oslo, PO Box 1080, Blindern, 0316 Oslo, Norway, [4]Centre for Cancer Biomedicine, Faculty of Medicine, University of Oslo, PO Box 4950, Nydalen, 0424 Oslo, Norway, [5]Statistics for Innovation, Norwegian Computing Center, 0314 Oslo, Norway and [6]Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, PO Box 4950, Nydalen, N-0424 Oslo, Norway

Received August 24, 2012; Revised March 7, 2013; Accepted March 12, 2013

ABSTRACT

The study of chromatin 3D structure has recently gained much focus owing to novel techniques for detecting genome-wide chromatin contacts using next-generation sequencing. A deeper understanding of the architecture of the DNA inside the nucleus is crucial for gaining insight into fundamental processes such as transcriptional regulation, genome dynamics and genome stability. Chromatin conformation capture-based methods, such as Hi-C and ChIA-PET, are now paving the way for routine genome-wide studies of chromatin 3D structure in a range of organisms and tissues. However, appropriate methods for analyzing such data are lacking. Here, we propose a hypothesis test and an enrichment score of 3D co-localization of genomic elements that handles intra- or interchromosomal interactions, both separately and jointly, and that adjusts for biases caused by structural dependencies in the 3D data. We show that maintaining structural properties during resampling is essential to obtain valid estimation of *P*-values. We apply the method on chromatin states and a set of mutated regions in leukemia cells, and find significant co-localization of these elements, with varying enrichment scores, supporting the role of chromatin 3D structure in shaping the landscape of somatic mutations in cancer.

INTRODUCTION

The spatial organization of chromatin is of major importance to key processes in the cell. Recently, several studies have shown that, in addition to regulatory functions (1,2), long-range DNA interactions are associated with the mutational landscape and chromosomal alterations in cancer genomes (3–5). Therefore, understanding how DNA is organized in the nucleus is crucial.

One recently published technique called Hi-C (6), has been shown to successfully map genome-wide 3D interactions in several species (7–9). Briefly, the Hi-C method uses formaldehyde to cross-link the DNA, which is subsequently digested using a restriction enzyme, and then paired-end next-generation sequencing determines the frequency of interactions between all pairs of restriction fragments. Other techniques based on chromosome conformation capture (10) include 5C (11) and ChIA-PET (12).

Despite these recent breakthroughs in experimental techniques for mapping chromatin 3D interactions, few tools have been developed to handle the large amounts of data that are produced in a statistically sound way.

We are interested in evaluating whether a set of regions in the genome (our 'query set of interest') are spatially closer to each other than what would be expected by chance. The Hi-C data will typically consist of restriction fragments that can be concatenated into bins of a certain constant size, which we will call genomic elements. We wish to evaluate whether a predefined subset of these elements has significantly higher interaction frequencies than what would be expected by chance. As is obvious, both the choice of query set and what we mean by chance is crucial to this question.

*To whom correspondence should be addressed. Tel: +47 2278 1778; Fax: +47 2222 421; Email: ehovig@radium.uio.no

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

One of the first computational methods to handle this question was proposed by Botta *et al.* (13). In this study, they assumed a null hypothesis where interactions were considered as independent and could therefore be randomized independently, using uniform resampling. They then compared the number of observed interactions with the average number of interactions in the randomized samples. Similarly, Duan *et al.* (7) and Dai and Dai (14) suggested that the number of interchromosomal interactions within a set of genes is hypergeometrically distributed, based on the assumption that the interactions are independent. However, as the 3D structure implies both transitive relations (if i is close to j and k, then j and k are also close) and correlation between certain pairs of interactions, these independence properties are not valid.

The dependency between interactions was recently pointed out in an article by Witten and Noble (15). In the same article, the authors proposed a simple resampling-based method for evaluating the overrepresentation of interchromosomal interactions in a set of genomic elements. They considered interaction frequencies in binary form, where true interactions were defined as interchromosomal interactions at a false discovery rate <0.01. Letting the size of the query set be n, they uniformly drew n new elements from the total population, and compared the number of interactions in the randomly chosen set with the number of interactions in the original set, keeping the number of elements on each chromosome constant. In this way, they obtained an estimate of the P-value according to the null hypothesis that the set of interest shows no more co-localization than a randomly chosen set of elements. Using this resampling approach, the global 3D structure is maintained, and therefore also the transitive properties. However, the dependency between interactions close in sequence is not preserved.

The Witten and Noble (15) method is designed to work only for interchromosomal interactions, and therefore abundant cis-acting interactions cannot be assessed. Additional properties will have to be considered when taking into account intrachromosomal interactions. Random close contacts in the DNA molecule cause systematically higher numbers of interactions for regions close in sequence compared with more distant regions. Such effects need to be adjusted for when testing on interaction frequencies within a chromosome.

There are several properties of the query set of interest that can be important to preserve in a hypothesis test setting when considering the total data set. Examples of such properties are the proportion of genomic elements close to centromeres and telomeres, or the GC content in the query set of interest. We show in this article that ignoring such features may cause skewness in the P-value distribution under the null model. This is because the interaction frequencies have varying distributions throughout the genome.

Imakaev *et al.* (16) showed that the three first eigenvectors of the bias corrected Hi-C data capture global patterns of chromatin interactions. The authors showed enrichments of contacts between genomic regions with similar corresponding elements in the first eigenvector. Because this eigenvector is strongly correlated with GC content, it implies that regions with similar GC content have a higher chance of interacting than regions with different GC content. In addition, they showed that the second and third eigenvectors pick up patterns relating to the relative positioning along the chromosome arms, where centromeric and telomeric regions are enriched for contacts within these regions more than between. The first eigenvector is related to the two-compartment model, where chromatin is divided into open and closed compartments, proposed in (6). Here, they also reported higher correlations between interaction frequencies within compartments compared with between compartments.

In this article, we present a genome-wide hypothesis test for inter- and intrachromosomal interactions, either separately or jointly, that can take into account structural properties due to both sequence-based distance and varying compartmental structure defined as domains along the chromosomes. We evaluate the method on both simulated and real data, and find that it performs well in all circumstances. Software for these tests is available online.

MATERIALS AND METHODS

A genome-wide hypothesis test of 3D co-localization of genomic elements

Based on knowledge about the spatial organization of a genome, there are some distinct and important properties to be considered in a hypothesis test context.

We are more likely to observe intrachromosomal interactions between elements with low sequence-based distance along a chromosome compared with high sequence-based distance (see Figure 1a), as shown in Lieberman-Aiden *et al.* (6). Consequently, the expectation and variance of the interaction frequencies depend on the sequence-based distance. For interchromosomal interactions, the sequence-based distance is undefined, and therefore the expectation and variance are constant in this case. In the calculation of the test statistics, we adjust for the different expectations and variances of inter- and intrachromosomal interactions given their sequence-based distance.

To maintain the transitive properties (see Figure 1b), we will randomize the query region of interest instead of the 3D structure. Still, in such a randomization, we need to consider the dependency between the interaction frequencies.

We want to test if a set of genomic elements (our 'query set of interest') has a higher 3D co-localization than what would be expected by chance. Our hypotheses are as follows:

H_0 : The query set of interest has the same 3D co-localization as a random set,
H_1 : The query set of interest has more 3D co-localization than a random set.

What we mean by 'random set' can vary according to what structural properties of the query set we want to preserve, and will be specified later.

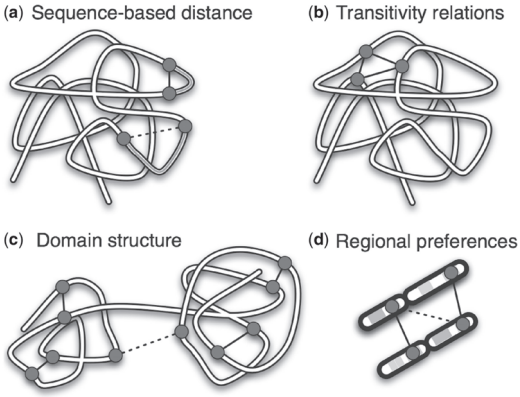

Figure 1. An overview of important structural features in chromatin 3D data, and how they are accounted for in the method. High and low interaction frequencies are shown as solid and dotted lines respectively, between selected genomic elements (circles). (**a**) Relationship between sequence-based distance (grey lines) and 3D contact frequency is corrected for using Equation 1. (**b**) All transitivity relations are preserved by randomizing the genomic elements only, and not the 3D interactions. (**c**) Interactions within domains are more prevalent than between domains. (**d**) Two genomic elements in the same relative position on the chromosome are more likely to interact than genomic elements on different positions. Both (c) and (d) are taken into account by using the domain randomization procedure. All these structural features lead to correlation between interactions with low sequence-based distance, which we take into account by using the CCD randomization procedure.

We will now describe a test statistic that measures the amount of 3D co-localization in the query set of interest. Let genomic element a_i be the element that starts on base pair i on chromosome a. Let $m_{a_ib_j}$ be the interaction frequency between genomic elements a_i and b_j. We calculate the sequence-based distance corresponding to an interaction as

$$\delta = \begin{cases} |j-i| & \text{if } a = b \\ \infty & \text{if } a \neq b. \end{cases}$$

If $a = b$, $m_{a_ib_j}$ corresponds to an intrachromosomal interaction, $\hat{E}(m|\delta = k)$ is the empirical mean of all intrachromosomal interaction frequencies with sequence-based distance $k = |j - i|$ and $\widehat{sd}(m|\delta = k)$ is the sample standard deviation. When $a \neq b$, $m_{a_ib_j}$ corresponds to an interchromosomal interaction, $\hat{E}(m|\delta = \infty)$ is the empirical mean of all interchromosomal interaction frequencies and $\widehat{sd}(m|\delta = \infty)$ is the sample standard deviation. If the number of observed interactions is low for certain high δ, it is advisable to assemble these into larger groups such that the estimation will be more accurate. Let $m^*_{a_ib_j}$ be the corrected interaction frequencies, which are adjusted for the expectation and standard deviation given δ like the following:

$$m^*_{a_ib_j} = \frac{m_{a_ib_j} - \hat{E}(m|\delta)}{\widehat{sd}(m|\delta)} \quad (1)$$

Let S_a^{int} be the set of base pairs corresponding to the genomic elements of interest on chromosome a, and let $Q = \bigcup S_a^{int}$ be our query set of interest over all chromosomes. The corresponding test statistic becomes the sum over all possible inter- and/or intrachromosomal corrected interaction frequencies m^* from Equation 1 in our query set Q:

$$t = \frac{1}{M} \sum_{a_i, b_j \in Q} m^*_{a_ib_j} \quad (2)$$

where M is the number of terms in the sum. Under the null hypothesis, the expected value of the numerator of Equation 1 will be close to zero, and therefore the test statistic in Equation 2 will be close to zero. We know that the variance of the test statistic is the sum over the variance for each corrected interaction frequency, plus the sum over the covariances between all pairs of corrected interaction frequencies. It follows that when the genomic elements in Q are close in sequence, the covariance between interaction frequencies increases, along with the variance of the test statistic.

We estimate the P-value using a permutation test, and resample R random sets. In the permutation of genomic elements, it is important to maintain the query set configuration, meaning the sequence-based distance between the genomic elements of interest. We choose to sample new positions by randomizing the order of the consecutive distances between the genomic elements in the query set. Thereby, the set of all successive distances between the elements in the query set are conserved (see Supplementary Figure S1). This leads us to the following Monte Carlo (MC) randomization strategy, which we name Conserved Consecutive Distances (CCD):

- Calculate t_{obs}, the test statistic from Equation 2 based on the query set of interest Q.
- Calculate sequence-based distance d_a between all pairs of consecutive genomic elements in S_a^{int} for all a.
- Repeat the following procedure for $r = 1, \cdots, R$.
  - For each chromosome a, let S_a^r be a random set, where the order of the sequence-based distance d_a is randomized. It follows that $|S_a^r| = |S_a^{int}|$.
  - Let t_r be the test statistic from Equation 2 based on the random set S_a^r for all a.
- We calculate the exact Monte Carlo P-value, described in (17)

$$p = \frac{\sum_{r=1}^{R} I(t_r \geq t_{\text{obs}}) + 1}{R + 1} \quad (3)$$

Testing for alternative hypotheses with lower co-localization, or testing for either lower or higher co-localization, is done in exactly the same way, but with a trivially modified P-value calculation.

We quantify the 3D co-localization of elements in the query set by calculating an enrichment score S. This is given as the ratio of the average observed over average expected co-localization. In the Supplementary methods, we provide a detailed description of the calculations.

When presented in percentage, we give the enrichment as $(S-1)100\%$.

The domain randomization procedure

It has recently become clear that the structural properties of mammalian chromatin are not constant throughout the entire genome, but varies locally depending on both GC-content and on relative positioning along the chromosome arms [see (16)]. The GC-dependent variation is related to the two-compartment model proposed in (6), where the nucleus is compartmentalized into open and closed chromatin. Therefore, in addition to conserving the consecutive distances within the query set during the randomization, it is often necessary to conserve these additional structural features as well (i.e. being more strict in the definition of a random set). In other words, we compare our query set with random sets with similar properties as the query set (see Figure 1c and d).

To conserve the structural features in the hypothesis test, we divide the genome into domains such that all genomic elements within the same domain have the same desired properties. We then use the CCD randomization strategy separately within each domain. Note that this strategy will not necessarily conserve the consecutive distances between adjacent domains. This, however, is not critical, as interactions are much more prevalent within than between domains.

For comparison, a 'global' randomization is performed in the form of using the CCD randomization strategy on the entire chromosome arm.

To evaluate the difference between the presented randomizations, we used two publicly available data sets and looked at two important properties to define the domains. First, we looked at the amount of genomic elements in open and closed compartments, we then considered the relative position of the genomic elements along the chromosome arm. We classified the genomic elements into open and closed compartments using the same method as Lieberman-Aiden *et al.* (6) (by looking at the sign of the first principal component). To categorize the position of the genomic elements on the chromosome, we divided the chromosome arms into six equally sized groups. To investigate the influence of the domain randomization, we chose 1000 query sets of size 50 at random (with the same domain properties), and compared the resulting *P*-values to the *P*-values when using the global randomization procedure. By definition, the *P*-values are uniform when using the domain randomization, but this does not need to be the case when using the global randomization.

Simulated data and method evaluation

To validate the CCD randomization strategy, we simulated 3D structures where H_0 was true by definition, and inspected the distribution of *P*-values for a large set of such structures. The *P*-values should be uniformly distributed if the resampling procedure is valid.

The 3D structures were simulated using random walks of size 500 inside a reflecting sphere. Two independent sequences (chromosomes) were simulated using the following algorithm $X_{a_i} = X_{a_{i-1}} + \frac{r_i}{||r_i||}$ for $i = 2, \ldots, 500$, where X_{a_1} was, for simulated chromosome a, a random starting 3D position sampled within a sphere with a diameter of $\sqrt{500}/2$. r_i was sampled from a 3D Gaussian distribution with $\mu = 0$ and $\sigma = 1$. Each step $X_{a_i} - X_{a_{i-1}}$ had length one and a random direction in the 3D space. The simulated interaction frequency was defined as $\log(1/||X_{a_i} - X_{b_j}|| + 1)$ for all possible paired genomic elements between and within the simulated chromosomes. We simulated 5000 such 3D structures containing two chromosomes each.

We compared our test statistic with an uncorrected version defined in the same way as Equation 2, except that we summed over 'uncorrected' interaction frequencies $m_{a_i b_j}$. We compared our Monte Carlo randomization strategy CCD with a simpler strategy where we resampled random sets S_a^r on each chromosome a, by sampling the same number of genomic elements uniformly distributed along the chromosome. We call this MC-strategy 'UNI'. In total, we compared four different approaches with increasing degree of sophistication: UNI with uncorrected test statistic, UNI with corrected test statistic, CCD with uncorrected test statistic and CCD with corrected test statistic. The distribution of interaction frequencies was similar over the entire simulated genome, so we did not need to use the domain randomization procedure in this particular test.

To show the effect of variations in the configuration of S_a^{int}, we evaluated all four approaches on three different types of S_a^{int} that were meant to represent a wide range of cases. The first type of query region consisted of 10 genomic elements uniformly sampled on each chromosome. In this case, we considered an ensemble of 150 query regions to see the distribution of the *P*-values in the average case. The second type of query region had 10 unique positions with high dispersion on each chromosome. Here, the positions were sampled using regularly spaced positions with added noise from a uniform distribution between 0 and 10. The last type was a set of 10 genomic elements heavily clustered on each chromosome. The positions for the genomic elements were sampled using ten unique positions from a Gaussian distribution with $\sigma = 10$ centered on the middle of the chromosome.

Specific versus regional co-localization

In analysis of real data, it is of interest to know whether a set of elements is co-localized simply because they are found in larger regions with general closeness, or if the query set itself is specifically co-localized compared with its near neighbors. Precise enrichment could potentially have a different interpretation than a more regional co-localization. To evaluate the specificity of the co-localization, we first perform a hypothesis test on the query set Q and calculate the *P*-value, and then subsequently perform hypothesis tests on neighboring query sets Q_k, where each genomic element is shifted k elements in a random direction from the original position in Q. We let $k \in (1, \cdots, K)$ and look at how fast the enrichment scores and the *P*-values change, according to k.

Software

All algorithms have been implemented in a publicly available statistical web toolkit called the Genomic Hyperbrowser, at http://hyperbrowser.uio.no/3d-coloc/ (18).

Publicly available data sets used

We used three different publicly available data sets with bin sizes varying from 100 kb to 1 Mb to evaluate inherent properties of chromatin 3D data and for hypothesis testing. All data sets used are adjusted for technical bias using the method of Imakaev *et al.* (16). For evaluating the domain randomization procedure, we used IMR90 and human embryonic stem cell (hESC) Hi-C data from (9). To test for co-localization of elements marked by somatic mutations, we used K562 Hi-C data from (6), and somatic mutations in leukemia patients from (19). We masked out centromeric, telomeric and gap regions, and performed the randomization within each chromosome arm separately.

RESULTS

In this section, we show how the dependency between interaction frequencies changes according to the sequence-based distance between the interactions, and use simulated data to validate the CCD randomization, which takes this dependency into account. With the publicly available data, we compare global and domain randomization, and use these methods to analyze the hypothesis that chromatin states are co-localized, and the hypothesis that mutated regions in leukemia patients are co-localized.

Interaction frequencies depend on sequence-based distance

A major motivation for the choice of randomization procedure is the occurrence of correlations between the interaction frequencies also after correcting for different sequence-based distances. We are strengthening this statement by showing that pairs of interactions with low sequence-based distance have similar corrected interaction frequencies. Specifically, we calculate the absolute difference $|m^*_{a_i b_j} - m^*_{a_k b_l}|$ between all pairs of contacts. For each intrachromosomal pair, we find their sequence-based distances defined according to the smallest distance $min(|i-k|,|j-l|,|j-k|,|i-l|)(\Delta_1)$ and the distance between the remaining two genomic elements (Δ_2). For instance, if $\Delta_1 = |i-k|$, then $\Delta_2 = |j-l|$, or if $\Delta_1 = |j-k|$, then $\Delta_2 = |i-l|$. For each interchromosomal interaction, Δ_1 is defined as $min(|i-k|,|j-l|)$ and Δ_2 to be $max(|i-k|,|j-l|)$. When Δ_1 is small, one genomic element from each of the two interactions has low sequence-based distance. When, in addition, Δ_2 is small, the other two genomic elements from each of the two interactions also have low sequence-based distance.

In Figure 2, we show the dependency between intrachromosomal interaction frequencies in hESC Hi-C data (9), measured by the average absolute difference, as explained above. As the figure shows, the corrected interaction frequencies tend to be more similar when both Δ_1 and Δ_2 are low, i.e. the interactions have low sequence-based distance. The interaction frequencies seem to be particularly similar for interactions that are separated by <5 bins on either end. This emphasizes the need to maintain the structure in the randomization for interactions with low sequence-based distance. We see the same trends for the IMR90 cell line (9) in Supplementary Figure S2. There does not seem to be a large difference between bin sizes, although the interaction frequencies are more similar when the bin size is large compared with when the bin size is small. This could be because the interaction distribution is smoother for higher bin sizes. We see similar trends for corrected interchromosomal interaction frequencies (data not shown). To maintain this structure, we have chosen to conserve consecutive distances during the randomization (i.e. using the CCD method). Figure 2 also shows that the dependency structure is similar for the random walk structures, even though these interaction frequencies are more similar overall owing to the lack of noise in these structures.

The resampling produces valid *P*-values

To validate the CCD resampling procedure, we looked at the distribution of the *P*-values in simulated data where H_0 was true. A valid procedure for *P*-value estimation

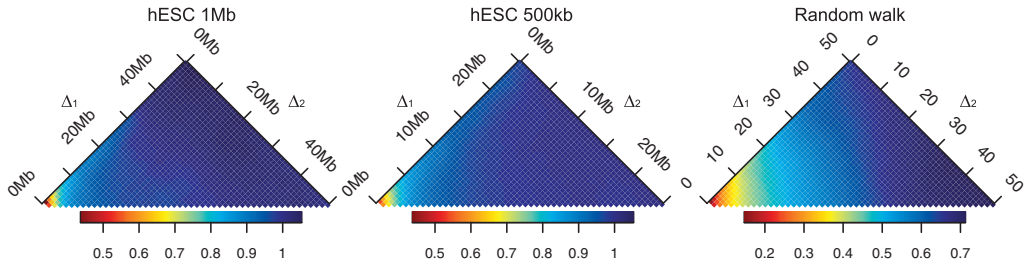

Figure 2. The average absolute difference $|m^*_{a_i a_j} - m^*_{a_k a_l}|$ between all pairs of corrected intrachromosomal interaction frequencies, given the two distances $\Delta_1 = min(|i-k|,|i-l|,|j-k|,|j-l|)$ and Δ_2 equal to the distance between the remaining two genomic elements. When Δ_1 is small, one genomic element from each of the two interactions have low sequence-based distance. When, in addition, Δ_2 is small, the other two genomic elements from each of the two interactions also have low sequence-based distance. The two sequence-based distances are given in million base pairs (Mb) along the genome. On the left, we see the result using hESC data (9) with bin size 1 Mb, in the middle, using a bin size of 500 kb and to the right, the simulated random walk structures.

should produce a uniform distribution of *P*-values under H_0. We simulate 5000 structures of two chromosomes under H_0 (see 'Materials and Methods' section). The distribution of the simulated interactions in the random walk structures are similar across the structure, so we use global randomization on the entire chromosomes.

Figure 3 shows the resulting *P*-value distributions for both simulated intra- and interchromosomal interactions considered jointly. As expected, both the corrected and uncorrected test statistics give uniformly distributed *P*-values when the genomic elements were in fact generated uniformly. In the middle panels, we see the distribution of the *P*-values in a more intricate case, i.e. when the query sets are spread out over each chromosome. With spread genomic elements, we observe small uncorrelated interaction frequencies. Using the uncorrected test statistic in combination with UNI, we obtain *P*-values shifted toward 1, as the true distribution of the uncorrected test statistic has lower expectation and variance than the UNI approximation. Choosing a clustered query set gives *P*-values that are biased in the other direction, as here, the true distribution of the uncorrelated test statistic has higher expectation and variance than the UNI approximation. The only satisfactory estimation of the *P*-value for all types of query sets is given by CCD in combination with the corrected test statistic, as we in this situation correct for both the expectation and the variance of the test statistic.

In Supplementary Figure S5, we see the resulting *P*-values for all combination of methods, namely UNI and CCD with both uncorrected and corrected test statistics. The same validations were also performed on simulated intra- and interchromosomal data separately (see, respectively, Supplementary Figures S3 and S4). The overall conclusion is the same: the only method that always gives uniformly distributed *P*-values for every considered query set when H_0 is true, is the corrected test statistic in combination with CCD. For the remainder of the analysis, we will exclusively use the CCD in combination with the corrected test statistic in Equation 2.

Taking into account domain structure is necessary for biologically meaningful *P*-values

As proposed earlier, it is also possible to use a more strict null hypothesis where we randomize within predefined domains. In this section, we use two Hi-C data sets from (9) to evaluate the domain randomization procedure. We conserve two important properties, first the amount of genomic elements within each open and closed compartment, second the relative positioning of the genomic elements along the chromosome arm, as explained in 'Materials and Methods' section. If the *P*-values are the same using both global and domain randomizations, then the interactions are equally distributed within the domains compared with the entire genome.

In Figure 4, we see, in the left panel, the *P*-values for query sets from the closed compartments, and in the right panel, query sets from the open compartments. In both cases, the *P*-values, when using the global randomization, tend to be close to zero. In other words, all the query sets in either open or closed compartments have larger 3D co-localization if we compare them with random sets from the entire genome. These results correspond with the findings in Lieberman-Aiden *et al.* (6) where they show that there are higher 3D contacts when the genomic elements are in the same compartments compared with when they are distributed between compartments.

In Figure 5, we see the *P*-values for query sets close to the telomeres (left panel), close to the center of the chromosome arms (middle panel) and close to the centromeres (right panel). Query sets in either end of the chromosome arm give *P*-values close to zero when using the global randomization, meaning they have larger 3D

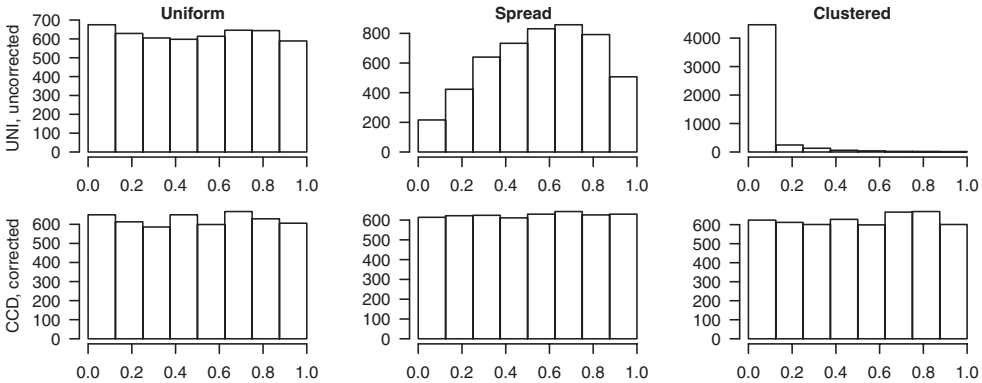

Figure 3. Each plot shows the histogram of 5000 *P*-values found by performing hypothesis testing on simulated 3D structures (based on a random walk procedure as explained in 'Materials and Methods' section) where our null hypothesis is true. The tests are performed on both intra- and interchromosomal interactions simultaneously. The upper row display the least complex approach, using the Monte Carlo resampling strategy UNI and the uncorrected test statistic. The bottom row shows the results from the Monte Carlo resampling strategy CCD and corrected test statistic from Equation 2. The three columns represent the three different configurations in the query sets of interest, uniformly distributed (left column), spread (middle column) and clustered (right column).

co-localization than if we compared them with random sets from the entire genome. In contrast, if we choose genomic elements in the middle of the chromosome arms, we find that they have lower 3D interaction, than if we compared them with random sets from the entire genome. This is similar to the results reported in Imakaev *et al.* (16).

When analyzing genomic elements all located in telomere, centromere or open compartments, one should avoid using the global randomization, as this hypothesis test will always give significance, as seen in the right plot in Figure 4.

In Supplementary Figures S6–S8 we see the results of the same test, using bin sizes 1 Mb, 500 kb and 200 kb, respectively. The analyses are performed on interactions from intrachromosomal, interchromosomal and both combined. In Supplementary Figures S9–S11, we see the results of the same analyses on cell line IMR90. In some of the cases, for example, when the genomic elements in the query set are close to the centromere, there are different results when comparing intra- and interchromosomal interactions. This is reasonable because intra- and interchromosomal interactions potentially represent very different features.

3D co-localization correlates with chromatin state activity

We have demonstrated that our method is capable of producing uniformly distributed *P*-values under H_0 (see Figure 3). However, it is also of interest to confirm that the method produces significant *P*-values when H_0 is not true. We performed a genome-wide test of co-localization for three different sets of genomic elements defined according to chromatin state activity in human embryonic stem cells [using the chromatin states as defined in Ernst *et al.* (20)]. We therefore classified the 100 kb Hi-C bins in human embryonic stem cells (9) into three categories: All bins covered by 'active promoter', all bins covered by 'strong enhancer' and all bins covered by >50% 'polycomb repressed' regions. For each of these three sets of genomic elements, we performed a hypothesis test using the global randomization and the domain randomization methods with two different domain classifications (open and closed compartments, and chromosome arm positions divided into six groups). All tests were performed on intra- and interchromosomal interactions separately, in addition to jointly. In Figure 6, we see the *P*-values and enrichment scores (see Supplementary Figures S12–S14 for test statistic distributions). As expected, the regions marked by promoter or enhancer are significantly co-localized ($P \leq 0.001$), even after taking the domain properties of the query set into account. The enrichment scores represent average changes over the entire query set, thus for large query sets (like these) the values are generally low, and must not be confused with the traditional fold change in gene expression, where genes are analyzed individually. For both query sets, we see a decrease in enrichment score when comparing the global with the domain randomizations. Both enhancers and promoters are highly present in open compartments, making a global randomization problematic. This illustrates the importance of maintaining domain properties during randomization. Regions marked by Polycomb repressed states do not give significant co-localization, despite suggestions that Polycomb group proteins create silencing hubs (21). This could be due to the fact that relatively few Hi-C bins in this data set are spanned largely by Polycomb repressed regions, or that the regions are repressed in other ways than through chromatin interactions.

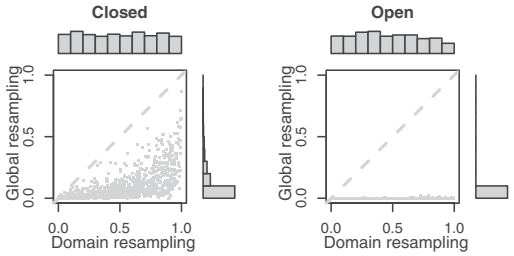

Figure 4. Evaluating random query sets with genomic elements in the closed (left panel) or open compartments (right panel). On the *x*-axis we see the *P*-values using the domain randomization, and on the *y*-axis we see the *P*-values using the global randomization. The results are based on both inter- and intrachromosomal interactions using the hESC data (9) with bin size 1 Mb.

Mutated regions in leukemia cells show statistically significant co-localization within chromosomes

Chromatin architecture increasingly appears to be of fundamental importance in many cancer-related processes. A recent study has suggested that somatic cancer mutation rates are largely influenced by chromatin organization (5). In that article, the authors showed that several heterochromatin-related epigenetic marks correlate positively with the frequency of somatic mutations in several cancers.

To gain further insight into the overall spatial patterns of mutated regions, we performed a genome-wide test of 3D co-localization of somatic mutations in leukemia samples (19) using a Hi-C data set from a human leukemia cell line (6), with bin sizes ranging from 100 kb to 1 Mb. For bin sizes 100 kb/200 kb, 500 kb and 1 Mb, bins were classified as mutated if they had at least one, two or three mutations within them, respectively. We then used the global randomization and the domain randomization methods with two different domain classifications (two compartments, and chromosome arm position based on six groups). All tests were performed both on intra- and interchromosomal interactions separately, and jointly.

In Figure 7, we see the *P*-value and enrichment score, and in Supplementary Figures S15–S18, we see the distribution of the test statistics under H_0. The top enrichment scores accompany the lower *P*-values, as expected. For intrachromosomal interactions, we have significant *P*-values, together with enrichment scores $\sim 2\%$. Such low enrichment scores, accompanied by significant *P*-values, indicate that either a small subset of the interactions have a large contribution, or that all interactions

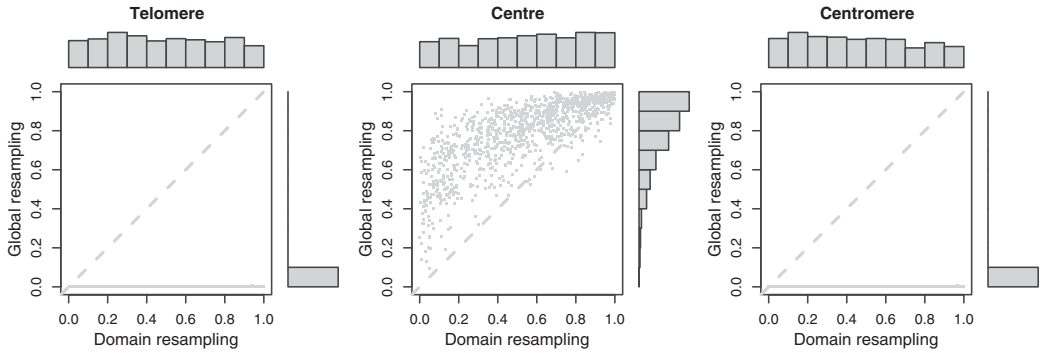

Figure 5. Evaluating random query sets with genomic elements close to the telomeres (left panel), close to the center of the chromosome arms (middle panel) or close to the centromeres (right panel). On the *x*-axis we see the *P*-values using the domain randomization, and the *y*-axis shows the *P*-values using the global randomization. The results are based on both inter- and intrachromosomal interactions using the hESC data (9) with bin size 1 Mb.

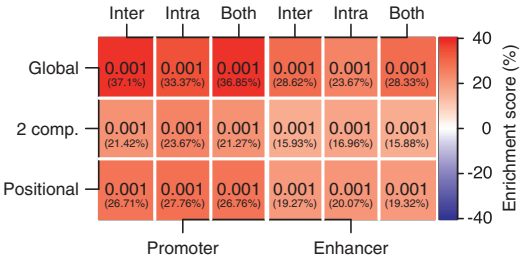

Figure 6. *P*-values and enrichment scores (in parenthesis) after testing on regions containing promoters (left) and enhancers (right) in hESC cells using 100 kb bins where we randomize globally (Global), within open and closed compartments (2 comp.), and within regions by dividing chromosome arms into six groups (Positional). Analysis was done on intra- and/or interchromosomal contacts.

contribute slightly to the significance. To get a better insight into the individual contributions of interactions in the query set, it is possible to look at a heat map over all individual test statistic terms for each interaction (see Supplementary Figures S15–S18). In our case, it seems that only a subset of interactions contribute to the enrichment.

It is interesting to note the difference in enrichment score when taking the domain structure of the query set into account in the randomization. We know that the query set initially is enriched in heterochromatin regions [as shown in (5)], which has lower co-localization compared with non-heterochromatin regions. As a result of this, the global randomization will place elements into open regions with generally higher co-localization. The domain randomization procedure will maintain the structural properties of the original query set, and will result in a more realistic enrichment score.

The reason why intrachromosomal interactions show a statistically significant enrichment could be owing to replication timing-related processes, as recently shown in (22). Here, they showed that the mutational landscape differ in early and late replication regions, with higher mutation frequencies in late replication regions. They also found that regions with similar mutational frequencies were close in 3D inside the nucleus. We also note that the observed co-localization could arise owing to reduced access of the repair machinery at inaccessible heterochromatic regions (23), or the increased exposure of mutagens in peripheral parts of the nucleus, causing mutations to cluster in specific regions of chromatin (24). If such clusterings of mutations are numerous and spatially separated in the nucleus, the 3D co-localization would mainly be enriched intrachromosomally, as the distance between clusters could be much larger than the distance within clusters. A consequence of this is low enrichment scores because interaction frequencies within clusters would typically be larger than its expected value, and interaction frequencies between clusters would typically be lower than its expected value.

The results emphasize the need for running tests at different resolutions, as *P*-values and enrichment scores can be radically different depending on the resolution chosen. We observe a trend toward lower *P*-values at lower resolutions, which probably can be attributed to reduced noise. The point at which the *P*-values stabilize could be the appropriate choice of bin size. We also note that any statistical test of 3D co-localization should be run on intra- and interchromosomal interactions both separately and jointly, as these could have different interpretations.

The specificity of 3D co-localization

To determine how specific the co-localization is, we performed a series of hypothesis tests where we shifted the elements in the query set away from their original positions. We did this by shifting each element in the query set in a random direction in steps varying from 1 up to 5 bins. In cases where a new position was invalid (typically for large k), we chose their position at random. Figure 8 shows the result of this analysis. The somatic mutations have low, but significant 3D co-localization, and we find significance in some of the query sets in the neighboring

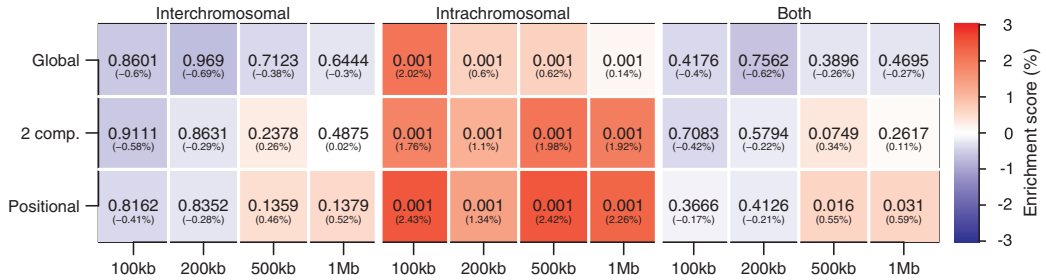

Figure 7. *P*-values resulting from hypothesis tests on the 3D co-localization of regions containing somatic mutations in leukemia cells. The colors and numbers in parentheses indicate the enrichment scores. Three different randomization strategies are used: the global randomization strategy (Global), domain randomization maintaining open and closed compartments (2 comp.) and domain randomization maintaining regional preferences by dividing chromosome arms into six groups. Analysis was done on intra- and interchromosomal contacts separately, and also jointly (Both). In addition, all tests were done on four different bin sizes (100 kb, 200 kb, 500 kb and 1 Mb) indicated at the bottom.

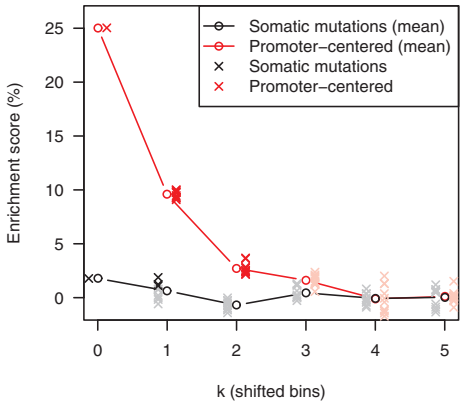

Figure 8. Estimated enrichment score after shifting each element in the original query set k bins in a random direction. The black solid line shows the average enrichment score of 10 independent runs on query sets defined by somatic mutations with a bin size of 100 kb, each indicated by a gray point (black point if the query set becomes significant at $\alpha = 0.01$). The red line shows a comparison with 10 independent runs on query sets defined by the promoter-centered elements (25) inside bins of size 100 kb in the same cell line (K562), where each run is either red (significant at $\alpha = 0.01$) or pink (non-significant).

bins $k = 1$, but when moving further away from the original query set, we lose both statistical significance and the quantified enrichment. A similar trend is seen for promoter-centered elements selected from (25), except for a steep drop in enrichment when shifting one bin, probably owing to the high specificity of promoter-centered interactions.

DISCUSSION

We have in this article addressed the important issue of dependencies between interaction frequencies in 3D data sets when estimating *P*-values in a hypothesis test context. We find strong dependency of interaction frequencies between contacts with low sequence-based distance (Figure 2), and show that such structures strongly affect the *P*-value estimation (Figure 3). We resolve such dependencies by using the CCD randomization strategy. We show that maintaining additional structural properties during randomization is necessary for biologically meaningful *P*-value estimation if the structures are not globally homogeneous. In mammalian genomes, for example, it was recently shown that interaction frequencies were highly dependent on GC-content and relative positioning along chromosome arms. We maintain such structure by randomizing within predefined domains, while simultaneously using the CCD randomization strategy. This article also presents methods for analyzing both intra- and intrerchromosomal interactons, separately and jointly. The results are presented with *P*-values and enrichment scores.

We have shown the importance of looking at both statistical significance and quantified contact enrichment, as significant *P*-values may be associated with different enrichments scores. Factors like sample size can modulate the *P*-value, meaning that larger query sets are more likely to be significant, given a signal. Given the low, yet significant, enrichments for intrachromosomal interactions between mutated elements in the leukemia cells, it is difficult to establish the biological meaningfulness of this result. Regardless, significance is found for all choices of bin size, and randomization methods. It can also be problematic to directly compare enrichment scores of functional interactions involving promoters and enhancers with 3D proximity of elements peripheral in the nucleus, as these can have different biological functions.

We have used a Monte Carlo strategy in the estimation of the *P*-value, as there is no adequate choice for the distribution of the test statistic. The main problem is to find a convincing distribution for all types of interaction frequencies that covers all the different aspects of the biological 3D structure. It is therefore highly important to critically evaluate the underlying null models and their relevant Monte Carlo options when using resampling methods in hypothesis testing to take into account the relevant structural properties. To do so, it is essential to know the data and their properties.

In principle, it could also be possible to randomize the 3D structure itself given that one could produce 3D structures from a valid null model universe. This, however, appears to be challenging, as a complete definition of a random chromatin structure needs to be established. We therefore emphasize that randomization in the query set, and not randomization of the 3D structure, is the natural resampling choice.

Our CCD randomization strategy only conserves the distance between successive genomic elements along the genome. This means that sequence-based distances between all possible pairs of genomic elements in the query set are not necessarily the same in the resampled set. It is, in theory, possible to maintain the entire structure in the query set with other choices of randomization procedures, for example, by randomly shifting the entire query set configuration along the genome. However, this leads to fewer resampling outcomes, and the resampling can rapidly become too constrained for useful analyses. We show that the relatively simple strategy of maintaining consecutive distances in the query set is sufficient to give correct *P*-values, at least in the query set configurations tested here. We also note that if we maintain the entire structure of our query set of interest in every Monte Carlo resampling, there would be no use of including the correction terms in the test statistic in Equation 2, as these would be constant across resamplings. However, we also show that this term is highly necessary when only consecutive distances are conserved.

In this article, we have looked at the question of co-localization between a set of genomic elements. Such co-localization is caused by spatial clustering of genomic elements in 3D, and is of interest in many settings. However, other interesting questions are not covered by this co-localization term, such as the 3D closeness between certain pairs of elements, or the comparison of 3D structures across treatments. We foresee that some of the same strategies as presented here probably will be valid in these settings as well. In practice, significant co-localization of a query set of interest is often resulting from a subset of the interaction frequencies. To visualize the query set consisting of mutated regions in K562, we clustered all elements according to intrachromosomal interaction frequency and visualized the resulting matrix as a heat map (see Supplementary Figures S19–S41). As the figures clearly show, only a subset of the elements seem to show enrichment of contacts. Therefore, a more specific test, such as co-localization of pairs of elements, would be able to find more detailed co-localizations. However, such a test would require more knowledge before running the test.

To evaluate the power of our method under various resampling constraints, we tested whether active parts of the genome were co-localized. We showed that active regions of the genome, such as promoters and enhancers, show significant and strong 3D co-localization, in contrast to polycomb repressed regions, which show no such enrichment. This holds true regardless of the resampling strategy used, which emphasizes the strong connection between genome function and structure.

While large consortia such as ENCODE (26) and the NIH Roadmap Epigenomics Program (27) have given a detailed annotation of epigenetic marks across several tissues and cell lines, the spatial interactions of these elements are not well understood. We believe rigorous statistical and computational methods, such as the one presented here, are needed to fill this gap.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–41 and Supplementary Methods.

ACKNOWLEDGEMENTS

We thank Sveinung Gundersen, Kai Trengereid and Tobias Gulbrandsen Waaler for helpful discussions in the initial phase of the project, and for help with the software implementation.

FUNDING

National Programme for Research in Functional Genomics (FUGE), and Statistics for Innovation (sfi^2), a centre for research-based innovation funded by the Norwegian Research Council. Funding for open access charge: The Norwegian Radium Hospital.

Conflict of interest statement. None declared.

REFERENCES

1. Kleinjan,D.A. and van Heyningen,V. (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.*, **76**, 8–32.
2. West,A.G. and Fraser,P. (2005) Remote control of gene transcription. *Hum. Mol. Genet.*, **14**, R101–R111.
3. De,S. and Michor,F. (2011) DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotechnol.*, **29**, 1103–1108.
4. Fudenberg,G., Getz,G., Meyerson,M. and Mirny,L.A. (2011) High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.*, **29**, 1109–1113.
5. Schuster-Böckler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, **488**, 504–507.
6. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
7. Duan,Z., Andronescu,M., Schutz,K., McIlwain,S., Kim,Y.J., Lee,C., Shendure,J., Fields,S., Blau,C.A. and Noble,W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
8. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Lebland,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-Dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, **148**, 458–472.
9. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
10. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
11. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. *et al.*

(2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
12. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.
13. Botta,M., Haider,S., Leung,I.X., Lio,P. and Mozziconacci,J. (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol. Syst. Biol.*, **6**, 426.
14. Dai,Z. and Dai,X. (2012) Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. *Nucleic Acids Res.*, **40**, 27–36.
15. Witten,D.M. and Noble,W.S. (2012) On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res.*, **40**, 3849–3855.
16. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
17. Phipson,B. and Smyth,G.K. (2010) Permutation *P*-values should never be zero: calculating exact *P*-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.*, **9**, Article 39.
18. Sandve,G., Gundersen,S., Rydbeck,H., Glad,I., Holden,L., Holden,M., Liestøl,K., Clancy,T., Ferkingstad,E., Johansen,M. *et al.* (2010) The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome biol.*, **11**, R121.
19. Puente,X.S., Pinyol,M., Quesada,V., Conde,L., Ordóñez,G.R., Villamor,N., Escaramis,G., Jares,P., Beá,S., González-Díaz,M. *et al.* (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **475**, 101–105.
20. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
21. Lanzuolo,C. and Orlando,V. (2012) Memories from the polycomb group proteins. *Annu. Rev. Genet.*, **46**, 561–589.
22. Liu,L., De,S. and Michor,F. (2013) DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. *Nat. Commun.*, **4**, 1–9.
23. Peterson,C.L. and Cote,J. (2004) Cellular machineries for chromosomal DNA repair. *Genes Dev.*, **18**, 602–616.
24. Misteli,T. (2007) Beyond the sequence: cellular organization of genome function. *Cell*, **128**, 787–800.
25. Li,G., Ruan,X., Auerbach,R.K., Sandhu,K.S., Zheng,M., Wang,P., Poh,H.M., Goh,Y., Lim,J., Zhang,J. *et al.* (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, **148**, 84–98.
26. ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.
27. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.

II

BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 11 2014, pages 1620–1622
doi:10.1093/bioinformatics/btu082

Genome analysis

Advance Access publication February 7, 2014

HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization

Jonas Paulsen[1,*], Geir Kjetil Sandve[2], Sveinung Gundersen[3], Tonje G. Lien[4], Kai Trengereid[5] and Eivind Hovig[1,2,3,*]

[1]Institute for Cancer Genetics and Informatics, Oslo University Hospital, PO Box 4950, Nydalen, 0424 Oslo, [2]Department of Informatics, University of Oslo, Problemveien 7, 0313 Oslo, [3]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, PO Box 4950, Nydalen, 0424 Oslo, [4]Department of Mathematics, University of Oslo, Problemveien 7, 0313 Oslo and [5]ELIXIR project, Department of Informatics, University of Oslo, Problemveien 7, 0313 Oslo, Norway

Associate Editor: Michael Brudno

ABSTRACT

Summary: Recently developed methods that couple next-generation sequencing with chromosome conformation capture-based techniques, such as Hi-C and ChIA-PET, allow for characterization of genome-wide chromatin 3D structure. Understanding the organization of chromatin in three dimensions is a crucial next step in the unraveling of global gene regulation, and methods for analyzing such data are needed. We have developed HiBrowse, a user-friendly web-tool consisting of a range of hypothesis-based and descriptive statistics, using realistic assumptions in null-models.

Availability and implementation: HiBrowse is supported by all major browsers, and is freely available at http://hyperbrowser.uio.no/3d. Software is implemented in Python, and source code is available for download by following instructions on the main site.

Contact: jonaspau@ifi.uio.no

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on October 18, 2013; revised on January 17, 2014; accepted on February 3, 2014

1 INTRODUCTION

Methods for detection of genome-wide chromatin 3D conformation, such as Hi-C (Lieberman-Aiden *et al*., 2009) and ChIA-PET (Fullwood *et al*., 2009), are drastically expanding our understanding of genome biology. However, statistical and computational methods to analyze chromatin conformation capture-based data are needed. Many of the available methods focus on data visualization, or are not suited for genome-wide statistical investigations (Baù *et al*., 2010; Servant *et al*., 2012; Thongjuea *et al*., 2013; Zhou *et al*., 2013). The structure of chromatin makes statistical analysis complicated, due to correlations between the interaction frequencies caused by both sequence-dependent and topological constraints (Paulsen *et al*., 2013). A few statistical tests have been proposed, with varying possibilities to account for structural dependencies (Botta *et al*., 2010; Kruse *et al*., 2013; Paulsen *et al*., 2013; Wang *et al*., 2013; Witten and Noble, 2012). Two useful command-line tools are the hiclib-package (Imakaev *et al*., 2012), and the HOMER software suit (Heinz *et al*., 2010), which both allow for noise-removal, outlier detection and compartment identification. The HOMER software additionally allows for identification of significant interactions in a given dataset, assuming a binomial distribution and a background model taking into account sequence-based and compartmental biases.

The global nature of these data allow for other types of statistical investigations beyond detecting significance of individual interactions. A common type of analysis is to analyze a set of genomic elements (genes, regulatory elements, transcription factors, etc.), and ask how this subset, or 'query track', is spatially arranged in 3D space as represented by a Hi-C dataset, for example. Here we present HiBrowse, a web-based analysis server for performing statistical analysis of 3D genomes in a range of different settings. The available statistics provide a flexible and expandable catalog of tools based on state-of-the-art statistical methods utilizing Monte Carlo (MC) and analytic methods as suited, in addition to a range of tools for visualization and hypothesis-generating investigations.

2 FEATURES AND METHODS

2.1 Data representation and analysis framework

We build on general software components of the Genomic HyperBrowser (Sandve *et al*., 2010, 2013), a web-based analysis server for genome-scale data. The graphical user interface (GUI) is based on Galaxy (Goecks *et al*., 2010), a user-friendly point-and-click environment familiar to many researchers. All tracks are based on a representation of elements as mathematical objects, consisting of points, segments, functions and variants of these [see Gundersen *et al*. (2011) for an in-depth discussion]. Any given analysis can be performed on all chromosomes, specific chromosomes or selected sub-parts of chromosomes, depending on the needs.

In practice, an analysis is initiated by selecting one or more tracks either from the HyperBrowser repository, or from the user history. At least one of the selected tracks must be a Hi-C (3D) track, and the accompanying selected tracks (called 'query tracks') determine the types of statistical analyses that are possible, and therefore selectable in the system.

A range of publicly available 3D-datasets have been installed in the repository. Since it has been shown that Hi-C and similar data can contain systematic biases, all the available Hi-C datasets have been corrected

*To whom correspondence should be addressed.

for such biases using the method of Imakaev *et al.* (2012). Furthermore, a specialized tool has been developed to allow users to upload their own Hi-C data (or similar) into the history, even if the dataset itself does not conform to well-known formats. See Supplementary Table S1 for a list of already installed and pre-processed Hi-C datasets.

2.2 Overview of statistical methods

Statistical tools are divided into two broad categories: hypothesis tests and descriptive statistics. Hypothesis tests are both MC based and analytical. Due to the complex structure of chromatin conformation capture data, finding suited explicit null distributions is generally not possible (Paulsen *et al.*, 2013; Witten and Noble, 2012), and even randomization of the data through MC is difficult. Therefore, we consistently perform permutations on the query track only. The hypothesis tests can be divided into three types, defined by the query track type, as illustrated in Figure 1A. For example, Points (P) are used to analyze general (all-versus-all) 3D co-localization by specifying a set of genomic elements using the BED format, while Linked Points (LP) are used to analyze 3D co-localization between selected pairs of elements by providing additional information about which genomic elements that should be linked together.

In the most basic case, if the user selects a set of points (genomic elements) in BED-format in addition to a Hi-C data track, one may ask whether all the genomic elements in the BED-file are more/less co-localized in 3D, in an all-versus-all fashion, than what would be expected by chance. In this case, the mean of the observed standardized interaction frequencies is compared to the expected value estimated from the permuted positions in representative regions of the rest of the Hi-C (3D) track. This analysis was introduced in Paulsen *et al.* (2013), and in this article we expand the methodologies by allowing a much wider variety of query tracks. For example, by specifying two point-tracks (two BED files), in addition to a Hi-C (or similar) track, the user can ask whether the points in track 1 are more/less co-localized with track 2, than expected by chance. In this type of statistical question, the permutations can be performed on both of the point-tracks, or by preserving one of the point-tracks completely.

It is also possible to specify particular interactions between a set of genomic elements, and compare these interactions with randomly permuted interactions within the same set of elements. In HiBrowse, interactions between genomic elements are defined using LP, a format described in detail elsewhere (Gundersen *et al.*, 2011). Such linked track types can easily be created by using a dedicated tool that converts from a simple BED file format containing information about which elements that should be linked together (see Supplementary Fig. S1, for an example). Since this type of analysis only permutes interactions intrinsically with regards to the query track, the positions of all elements will be completely preserved. This type of analysis should be used whenever specific interactions between genomic elements are considered, and it would be natural to compare with random links between the same elements. Since regions of the genome can have varying properties (active/inactive genes, open/closed chromatin, etc.), global shuffling of links between all selected elements is not always preferable. To take such properties into account during the permutation, each of the points can be marked by a value, such that the link-permutations will be performed by preserving the value-combinations on both sides of the links.

If the user wants full control over exactly what pairs of interactions that are allowed to take part in the link-permutations, it is possible to specify a case/control value on each of the links via a dedicated tool which accepts two BED files ('case' and 'control') of the same format as described above (see Supplementary Fig. S2, for an example). The case/control-linked elements can then be selected together with a Hi-C (3D) track, allowing the user to compare the interaction frequency of all the links marked as 'case' with the expected interaction frequency given by permuting the case/control labels. This type of statistic is optimal for data that is only sampled from a pre-defined set of elements of the genome, and where the user wants to find out whether a subset of these elements are co-localized in 3D.

Finally, it is possible to find statistically significant differences between two Hi-C datasets, for example comparing treatments [as e.g. in

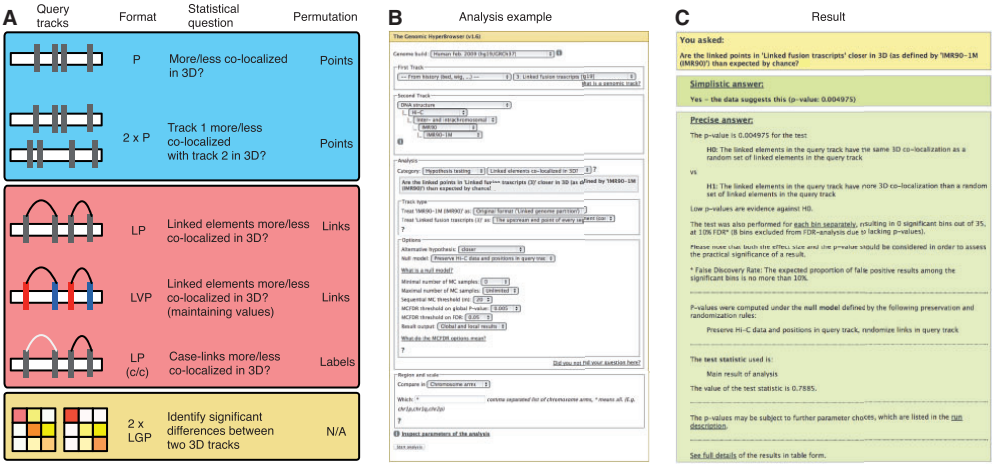

Fig. 1. (**A**) Overview of statistical hypothesis tests implemented in HiBrowse. See Gundersen *et al.* (2011) for an in-depth explanation of track types, and the Supplementary Material for details about each statistic. (**B**) Example of a HiBrowse analysis using the 'Linked elements more/less co-localized in 3D?' statistic, investigating whether fusion transcripts are co-localized in 3D. (**C**) Result page from the analysis, presenting the question asked by the user together with both a simplistic and a more detailed answer giving the *P*-value and model assumption details. Links are provided to full details of the results at individual chromosome regions

Rickman *et al*. (2012)]. The statistical test implemented for this type of analysis is based on the edgeR-tool (Robinson *et al*., 2010). Details about the mathematical formulation of the different types of statistics and their corresponding null-hypotheses are found in the Supplementary Material.

In addition to hypothesis tests, a range of descriptive statistics have been implemented. For example, each hypothesis test is accompanied by an enrichment score, giving the degree of over/under-representation of 3D co-localization, compared to the expected 3D co-localization (see Supplementary Material for details). Other types of available descriptive statistics are visualization of clustered Hi-C matrices as heatmaps or graphs, principal component analysis on Hi-C matrices and other summary statistics (see Supplementary Table S2 for a comprehensive list). All available analyses are described thoroughly on the help pages linked from the main site, where example histories are provided such that users can explore each statistic in detail. Demo-buttons are provided for all tools, giving small example runs. See Figure 1B and C for an analysis example.

Funding: This work was supported by the Norwegian Cancer Society [PR-2006-0433].

Conflict of Interest: none declared.

REFERENCES

Baù,D. *et al*. (2010) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol*., **18**, 107–114.

Botta,M. *et al*. (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol. Syst. Biol*, **6**.

Fullwood,M.J. *et al*. (2009) An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, **462**, 58–64.

Goecks,J. *et al*. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*., **11**, R86.

Gundersen,S. *et al*. (2011) Identifying elemental genomic track types and representing them uniformly. *BMC Bioinformatics*, **12**, 494.

Heinz,S. *et al*. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

Imakaev,M. *et al*. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

Kruse,K. *et al*. (2013) A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. *Nucleic Acids Res*., **41**, 701–710.

Lieberman-Aiden,E. *et al*. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Paulsen,J. *et al*. (2013) Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Res*., **41**, 5164–5174.

Rickman,D.S. *et al*. (2012) Oncogene-mediated alterations in chromatin conformation. *Proc. Natl Acad. Sci. USA*, **109**, 9083–9088.

Robinson,M.D. *et al*. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Sandve,G.K. *et al*. (2010) The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biol*., **11**, R121.

Sandve,G.K. *et al*. (2013) The Genomic HyperBrowser: an analysis web server for genome-scale data. *Nucleic Acids Res*., **41**, W133–W141.

Servant,N. *et al*. (2012) HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics*, **28**, 2843–2844.

Thongjuea,S. *et al*. (2013) r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data. *Nucleic Acids Res*., **41**, e132.

Wang,H. *et al*. (2013) Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin. In: *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. ACM, Washington, DC, USA, p. 306.

Witten,D.M. and Noble,W.S. (2012) On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res*., **40**, 3849–3855.

Zhou,X. *et al*. (2013) Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat. Methods*, **10**, 375–376.

III